Understanding Machine Learning – A Philosophical Inquiry of its Technical Lineage and Speculative Future

by

Felix Tun Han Lo

M.A., Simon Fraser University, 2016

M.Eng., Massachusetts Institute of Technology, 1996

B.Sc., Massachusetts Institute of Technology, 1995

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Communication
Faculty of Communication, Art and Technology

SIMON FRASER UNIVERSITY

Summer 2024

Declaration of Committee

Name:	**Felix Tun Han Lo**
Degree:	**Doctor of Philosophy (Communication)**
Title:	**Understanding Machine Learning – A Philosophical Inquiry of its Technical Lineage and Speculative Future**
Committee:	**Chair: Richard Smith** Professor, Communication **Frederik Lesage** Supervisor Associate Professor, Communication **Andrew Feenberg** Committee Member Professor Emeritus, Communication **Yuk Hui** Committee Member Professor, Erasmus School of Philosophy Erasmus University Rotterdam **Wendy Chun** Examiner Professor, Communication **Warren Sack** External Examiner Professor, Film & Digital Media University of California, Santa Cruz

Abstract

This dissertation presents a philosophical critique of machine learning based on an investigation into its technical lineage. It begins by explicating Martin Heidegger's remarks that cybernetics would take the place of philosophy and that "only a god can save us" from a technocratic society. But whereas Heidegger's critique assumes the universality of cybernetics, this assumption can be challenged by examining the transactions of the Cybernetics Conference. Such an examination exposes the inherent conflicts between disciplinary knowledge, which helps explain the failure of cybernetics in attaining scientific achievements. This dissertation further argues, even though cybernetics has shaped the historical development of computer science, artificial intelligence (AI), and machine learning, cybernetics and universal computing can be distinguished as two mutually imbricated intellectual traditions. AI research explores how human intelligence can be simulated on a universal computer, departing from the cybernetic objective of understanding the mechanisms of living organisms. In particular, Ray Solomonoff showed how his abstract machine-learning algorithm can recognize any subtle data patterns, anticipating the capability and limitation of deep learning.

Deep learning has made possible generative AI applications such as DeepBach, raising questions about the possibility of computational creativity or emotivity. This dissertation deliberates such questions by turning to Gilbert Simondon, whose model of philosophy derives from subatomic quantum behavior and other modern scientific theories, as opposed to everyday intuition applied to large objects. While cybernetics has also been influential to his philosophy, Simondon rejects the cybernetic mechanization of the living and its blurring of the life-machine boundary. Rather than conflating the human and the machine, Simondon's theories of concretization and individuation of transindividual relations suggest how technology co-evolves with the human and the social. These theories were adopted by Andrew Feenberg in his critique of the Internet and by Bernard Stiegler in his critique of algorithmic governmentality. Even though Feenberg's critique emphasizes the openness of the Internet while Stiegler's reveals the closed character of the 24/7 computational infrastructure, their interpretations of Simondon are compatible, as both recognize the revolutionary potentiality in human-technology co-evolution, which can be differentiated from J. C. R. Licklider's human-computer symbiosis.

Keywords: AI; Machine Learning; Solomonoff; individuation; potentiality; symbiosis

Dedication

To Teresa and Joanna, Mom and Dad, Ada, Eileen, and Adrian for their loving support.

To my Lord Jesus Christ for granting me the intellect and the perseverance to see this work to its completion.

Acknowledgements

This dissertation has been more than seven years in the making. First and foremost, I owe an enormous debt of gratitude to my supervisory committee. In particular, I feel super fortunate to have Frederik Lesage as my supervisor, who provides valuable guidance on my research direction while allowing me to freely explore themes and ideas that interest me the most. He is always available when I need any advice and has been an encouraging presence throughout this intellectual journey. I am also deeply indebted to Andrew Feenberg for shaping my intellectual growth under his mentorship. He has a knack for putting difficult concepts in ways that are easy to understand, and I will forever cherish the privilege of walking over to his office from the ACT lab for a long chat. I am also extremely grateful to Yuk Hui for willing to be on my committee in spite of his busy schedule and for his challenges to enhance the rigor of my arguments in this work.

I would also like to extend my heartfelt gratitude to all the people I met on my academic journey at Simon Fraser University: To Jason Congdon for providing invaluable institutional support; to other CMNS grad students, including Dylan Chandler, Matthew Greaves, Taeyoung Kim, Alberto Lusoli, and particularly the late Peter Zuurbier, for tips about the nuances of being a graduate student, for helping me edit my writing, and for the many conversations in our office or over lunches and coffees; to all the professors with whom I had the opportunity to attend classes—Gary McCarron, Jan Marontate, Peter Chow-White, Zoë Druick, and Rick Gruneau—as they have all left a lasting impact on my writing and thinking practices.

A significant aspect of this dissertation depends on an up-to-date awareness of the latest development in AI. In this regard, I would like to thank old friends from M.I.T. and from the industry—Patrick Chan, David Ma, David Maw, Stephen Siu, and Bernard Wong—for keeping me in the loop on the news along with other writings on AI, and for shedding light on the perspective of AI practitioners. I am also grateful to Wendy Chun and Matt Canute for inviting me to join the data science team in Digital Democracies Institute, where I gain exposure to new ideas and technologies in the interdisciplinary research on AI and humanities.

Table of Contents

List of Figures

List of Acronyms

AI	Artificial Intelligence
AGI	Artificial General Intelligence
GAN	Generative Adversarial Network
GPT	Generative Pre-trained Transformer
ILNFI	*L'individuation à la lumière des notions de forme et d'information*
LSTM	Long Short-Term Memory
MEOT	*Du mode d'existence des objets techniques*
RNN	Recurrent Neural Network
STS	Science and Technology Studies
UTM	Universal Turing Machine
XAI	Explainable Artificial Intelligence
XR	Extended Reality

Introduction

Background

Popular imagination on the potentials in artificial intelligence (AI) has made a resurgence since the turn of the millennium. When AI made its first splash in the 1950s and 1960s, constructing machines with intellect and thoughts appeared to be genuinely possible, and such possibility struck a chord with the dystopian fear that societies would one day be taken over by machines. This dystopian theme concerning AI was exploited in widely popular science fictions such as *I, Robot* (Asimov, 1950/2008), *2001: A Space Odyssey* (Kubrick, 1968), or *Do Androids Dream of Electric Sheep? (Dick, 1968/1972)*[1]. Then came the "AI winter" in the 1980s, a decade of futility with relatively little technical progress in AI research. The excitement and fears associated with AI dwindled along with the lack of significant inventions in this period. But with the breakthrough of deep learning over the past couple of decades, AI can now perform, and is potentially capable of performing, surprisingly intelligent functions that were deemed impossible in the past. Heated discussion about AI's vast implications has become especially more pronounced since the launch of OpenAI's ChatGPT. This AI chabot, implemented by fine-tuning OpenAI's generative pre-trained transformer (GPT) AI model, has stunned its users with its human-like and thorough responses, provoking various worries about its disruption on the job markets (Schell, 2023), on education (Heaven, 2023), and on new forms of societal threats (e.g., Chow & Perrigo, 2023; Elias, 2023a, 2023b; Harris & Raskin, 2023; Joseph Raczynski, 2023; Perrigo, 2023; Yudkowsky, 2023). This past decade of AI development has led to both excitement and fears over what unfettered development of AI technology may bring.

Foremost among the skeptical concerns are two dystopian visions of our future, "digital totalitarianism" and "technological singularity." According to the Israeli historian and philosopher Yuval Noah Harari, the threats of a "global useless class" and of "digital totalitarianism" are on the immediate horizon (Harari, 2018b). Harari sees AI as a major factor in disturbing the global balance. Immense wealth will be held in a few high-tech hubs, and the capitalist exploitation of labor in the 20th century will transit into the

[1] *Do Androids Dream of Electric Sheep?* was turned into the film *Blade Runner* (R. Scott, 1982).

creation of a class of people considered irrelevant in the new world order, the "global useless class." An even more ominous picture is a society under a digital form of dictatorship enabled by AI, a "digital dictatorship" that presents a serious threat to the liberal democracy (2018a, p. 46). To Harari, liberal democracy used to outperform dictatorship in the twentieth century because it is more efficient to process information and make decisions in a flexible and distributed economy that emphasizes free information flow. But recent development in AI "swings the pendulum" in favor of dictatorship from liberal democracy, because it is now more technologically efficient to host all data in a centralized site in order to train better algorithms. The future technological roadmap of combining artificial intelligence with biotechnology further gives humanity the power to reshape and reengineer life. Leaving such power in the hands of dictatorships could put future lives under total surveillance like in *1984* (Orwell, 1949). Or like *Brave New World* (Huxley, 1932), a government state may reshape individuals' genetic profiling to fit state planning. Not only will AI strengthen a government's totalitarian agenda, new AI tools for surveillance and for manipulating popular sentiments will make possible new ways for a dictatorship to exert controls over its citizens.

At the same time, the notion of "technological singularity" has gained much currency in recent years. The first articulation of the idea may have come from Ray Solomonoff's six milestones for AI (1985), in which the last milestone is a machine with many times the intellectual capacity of human societies. This idea of machinic superintelligence is later taken up by others (Joy, 2000; Kurzweil, 2000, 2005; Vinge, 1993), who argued that once computers achieve human-like intelligence, they will immediately surpass and become infinitely more powerful than human intelligence combined. The point in time when AI's intelligence surpasses that of humankind is called "technological singularity." The late renowned theoretical physicist Stephen Hawking, who developed a theory of space-time singularity to explain the origin of the universe, seemed to concur that "[t]he development of full artificial intelligence could spell the end of the human race" (Cellan-Jones, 2014). The threat that AI will one day become a superior "species" to the human race have stirred up popular imagination in recent years, as evident in the widespread success of the *Westworld* (2016—), *Her* (2013) and *Ex Machina* (2014). These media productions invite viewers to ponder whether a biotechnological evolution can indeed overtake the human species and put them into extinction. But not all experts see the same threat. Noam Chomsky expressed his dismay in an interview with *Singularity Webblog* (*Noam Chomsky on AI*, 2013). The podcast host asked Chomsky to give his opinion on "technological singularity," and he bluntly retorted that singularity is just science fiction. To him, we are still "eons away" from seeing AI automatons with emotions, creativity, and conscious wills.

But with the launching of AI agents like ChatGPT, which can communicate in human languages and pretend to be a human to send commands to other humans or machines via online platforms[2], the warning that AI could ultimately pose an existential threat to humankind may be closer in reality than many had previously thought. Some critics are claiming that "half of AI researchers give AI at least a 10% chance of causing human extinction" (Harris & Raskin, 2023; Tegmark, 2023).[3] The existential question has less to do with whether AI can have its own will and consciousness than the society's unpreparedness in safeguarding the mass deployment of AI automatons whose behaviors may not align with human interests and moral values. Geoffrey Hinton, considered by some as "the god father of artificial intelligence" due to his pioneering works in deep learning and the training of language models, quit his job at Google to warn the world about the existential risk that large language models may bring (Elias, 2023b). The argument goes, once the technology has mastered the human language and can reason linguistically like humans, autonomous AI agents can do anything a human can in the online world (Joseph Raczynski, 2023). Their actions may align with any malicious goals specified by their creators, or they may self-train and develop in an autonomous manner with no guarantee of aligning with our moral values and social norms (See *As AutoGPT Released, Should We Be Worried about AI?*, 2023). Misaligned AI behaves like buggy AI software in computer games, except now the possibilities of actions and the rules constraining them are practically no different from those facing real humans in their online representations. Such malicious or out-of-controlled automatons

[2] Rather than engaging purely in exchanges of text messages, a program like AutoGPT can write a python program and then executes the program. If the program is given root privilege, it can even download software and create new users to perform different actions. This is how an AI agent can act like an avatar representing a human user in the online world.

[3] This claim is based on the "2022 Expert Survey on Progress in AI" (Grace & Weinstein-Raun, 2022), but the figure is a bit misleading. Only 738 of 4271 responded to the survey, and only 162 of 738 answered the specific question (Sundin, 2023). Therefore, the survey result is not statistical relevant. The references illustrate rather the level of media hype about the existential risk of AI.

could pose existential threats to the human society because they are capable of replicating themselves into many copies, hacking into military computer servers to control weapons remotely, as well as coming up with and executing action plans that manipulate human activities. In this argument on existential threats, AI does not need to have a conscious will. It simply appears to have a conscious will by going out-of-control. This existential threat is thus closer in nature to the threats of nuclear weaponry than to the threat of technological singularity.

Facing such threats, Google CEO Sundar Pichai admonished that our society must quickly stipulate regulations to make AI safe for the world and align with human values. But we appear unprepared due to the mismatch between “the pace at which we can think and adapt as societal institutions [and] the pace at which the technology is evolving” (Elias, 2023a). Elon Musk, Steve Wozniak and dozens of academics called for an immediate pause in training “experiments” connected to large language models that are “more powerful than GPT-4” (Elias, 2023b; “Pause Giant AI Experiments,” 2023). Eliezer Yudkowsky, considered a founder in the field of aligning Artificial General Intelligence, goes further to suggest that we should “shut it all down” (Yudkowsky, 2023). But as Hinton remarked, stopping outright or even putting a brake on AI research is unlikely as a politically viable option, given how crucial AI is to a nation’s competitiveness on both the economy and the military front (Joseph Raczynski, 2023). In fact, there is no doubt that further advances in AI will bring many benefits to our society (Nadimpalli, 2017; Rossi, 2016; Yeasmin, 2019). Yet, there has never been such a moment in history in which AI has come to the forefront of human concerns, from preparing the youngsters for the unpredictability of future job markets to the potential instability of societal institutions across the globe, from the threat of a totalitarian army of AI agents to the real possibility of an existential threat to humankind posed by AI. People are suddenly bracing themselves for all kinds of survival and psychological challenges caused by AI.

At the same time, AI practitioners are well known for their penchant to anthropomorphize machines, to “talk about their artifacts using a vocabulary that is canonically used to describe human behavior” (Ekbia, 2008, p. 5). As evident in the

Talos of Greek mythology[4], Mary Shelley's *Frankenstein*, or Karel Čapek's *RUR-Rossum's Universal Robots* (1920),[5] the fascination with artificial life and with its dystopian possibility has been deeply ingrained in human culture. According to media theorist Simone Natale (2021), appealing to this fascination and deceiving users' perception are intrinsic characteristics of AI. Amidst all the anxious feelings and confusions, how can we discern between facts and fictions regarding AI? How can we adjudicate whether or not the latest warnings against AI's societal threats are just opiniated red herrings as some (e.g., Shingler, 2023) have contended? Can these threats be mitigated and kept in track by various technical and political measures? How can we tell whether all the recent hype about AI will soon fade, repeating the early history of AI, or if there are good reasons to project its transformative impact on our society?

These questions can only be properly answered if we have a better understanding of what AI is fundamentally about. My dissertation attempts to contribute to this fundamental understanding by conducting an immanent and philosophical critique of AI and machine learning. To attain an immanent understanding, I conduct an inquiry into the technical lineage of AI and machine learning, focusing on the intellectual activities during a particular thirty-year window of the post-war period. My inquiry includes the identification of ideas in cybernetics and universal computing that were instrumental to the development of AI and machine learning. Such ideas, which have shaped AI in the past, should continue to shape AI in the future. This approach makes possible a critique of AI without necessarily subscribing to the diverse opinions of AI experts or to the public perception of AI, which is easily swayed by the latest popular science fictions, the opinions of AI experts, or the trendiest AI inventions. I then step back to examine AI and machine learning from an epistemic distance, engaging in a dialogue with the critical thoughts of continental philosophers on cybernetics and universal computing. This dialogue brings to light the philosophies underpinning these technological and intellectual movements as well as the historical and social significance

[4] "Talos was a giant constructed of bronze who acted as guardian for the island of Crete. He would throw boulders at the ships of invaders, and would complete 3 circuits around the island's perimeter daily. According to pseudo-Apollodorus' Bibliotheke, Hephaestus forged Talos with the aid of a cyclops and presented the automaton as a gift to Minos" (*History of Artificial Intelligence - Wikipedia*, n.d.).

[5] This 1920 science-fiction play introduced the word "robot" to the science-fiction genre.

of these movements. After establishing the basis for understanding and analyzing AI and machine learning, I will look toward the not-so-distant future by alluding to science fictions based on realistic anticipation of impending AI technologies. I will evaluate and analyze these science fictions based on the previously established immanent and philosophical critique to deliberate on the social implication of AI in the not-so-distant future.

The Work's Organization

This dissertation is organized into nine chapters.

Chapter 1 presents the methodology of this dissertation. It begins with a discussion on the cultural gap between AI practitioners and humanities scholars by alluding to the animosity between early AI pioneers and the phenomenologist Hubert Dreyfus. My dissertation intends to move past this dichotomy between the technical culture and the humanities, combining a critical philosophical inquiry of AI with an immanent understanding about the potentials of machine learning. This immanent understanding can be attained by unraveling the interwoven strands of thought and clarifying the mutual influences between cybernetics, universal computing, AI and machine learning. This clarification can help elucidate the various threads of ideas associated with these fields. An investigation into computational theories can also contribute to this immanent understanding of AI. Of particular interest are the formal proofs that circumscribe the limits and potentials of machine learning, which is a subfield of AI responsible for all its recent advances. At the same time, for a critical inquiry of AI, a certain epistemic distance is needed. Hence my critique draws on continental philosophy more so than on analytic philosophy. This chapter further explains why, among the continental philosophers, Martin Heidegger and Gilbert Simondon are particularly significant for a social critique on AI technologies. My final analysis on the future of AI will be conducted based on this immanent understanding and critical philosophy, with the site of the analysis being science fictions written based on realistic projections of AI technology.

Chapter 2 attempts to explicate Heidegger's alarming claim that the emergence of cybernetics as a fundamental science in the mid-twentieth century marked the completion and the end of philosophy, and that only a god can save humanity from the resulting technocratic society. This dystopian view on the completion and the end of metaphysics in cybernetics can only be fully understood by reading his reflections on the historical development of the will of representational thinking, which culminates in the extreme form of Nietzsche's will to power as the technological will to will. This technological will to will is the essence of modern technology, which forces the calculation and arrangement of everything for technology itself. Heidegger's reflections on cybernetics and the technological will to will has been considered by some as a critique of techno-posthumanism and technological singularity. The epitome of a supreme technological will is artificial general intelligence with a conscious will of its own. But despite his remark that only a god can save us, Heidegger's later works mark his attempt to articulate how a salvation from modern technology may come about. According to Bret W. Davis in *Heidegger and the Will: On the Way to Gelassenheit* (2007), Heidegger went from the articulation of a "proper will" to a "non-willing" as a fundamental attunement of letting beings be, which serves as the alternative to the willfully positing of beings in cybernetic calculation and arrangement. For Heidegger, the extreme epoch at the end of the first beginning of the history of Western metaphysics is ironically the tipping point for an other beginning of the history of being marked by "letting beings be." While this eschatology of the transition from the epoch of the will cannot be "willed" by human beings, we may nonetheless participate in this transition between epochs. Heidegger's epochal eschatology seems particularly influential to Bernard Stiegler's critique of automatic societies.

Chapters 3 to 5 belong together. They examine AI from an immanent perspective by studying its history and its computational theories. In these chapters, I will explore the historical relations between the early history of AI, the cybernetics movement, and the early history of computing during the post-war period. In addition, I will explain the significance of computational theory on machine learning established in this period. Chapter 3 traces the genealogy of AI and challenges the commonly-held notion about the universality of cybernetics. Social critics, including Heidegger, characterize cybernetics as universal, as a fundamental science in its control of all beings. This characterization makes cybernetics an easy target for critique, but is far from conclusive if we directly examine the actual conversations that took place between the participants of the Macy Cybernetics Conferences. These conversations reveal the strife between conference participants due to incompatible assumptions held by different knowledge disciplines. The conferences and the movement were meagrely held together by themes such as negative feedbacks, homeostasis, or the blurring of boundary between humans and machines. All the participants share the desire to discover the mechanisms behind animal, human, and social behaviour and replicate these mechanisms in machines, but the various cybernetics themes cannot unify the diverse and conflictual scientific approaches between heterogeneous disciplines. They can only bring them into conversation. In the end, cybernetics ended in a historical failure with few lasting scientific or technical achievements. Nonetheless, bringing multiple disciplines into conversation did serve as the impetus for new research ideas. In this chapter, I take the stance that cybernetics and the computers as calculating machines belong to two distinct intellectual traditions that mutually influence each other. By intellectual tradition, I am referring to the evolving body of thoughts, ideas, and approaches that are passed down within the scientific community over time. Accordingly, the birth of AI can be attributed to the clash of ideas between the mainstream research of computing machines and the cybernetic idea of blurring the boundary between man and machines.

Chapter 4 continues this historical inquiry of AI by tracing the genealogy of machine learning. I argue that, in comparison to the rest of the discipline in AI and in computer sciences, machine learning inherits more directly from cybernetics. At the same time, machine learning, as a subfield of AI and computing, is distinguishable from the learning in cybernetic machines and in automata studies. This chapter depicts the electromechanical designs of cybernetic learning devices such as Claude Shannon's maze-solving mechanical mouse or W. Ross Ashby's Homeostat, and explains how these devices were designed to imitate the learning mechanisms in living organisms. In comparison, imitating living mechanisms has no place in machine-learning, which is concerned with abstract algorithms typically implemented in software. Like the rest of the field in AI and in computing, these abstract algorithms exploit the universality of a universal computer. I will argue how this prioritization of abstraction over imitation is evident in Alan Turing's papers on machine intelligence, in Arthur Samuel's checkers-playing machine, and in Ray Solomonoff's formal proofs on the limitations and potentials of machine learning.

Chapter 5 explains Solomonoff's formal proofs on algorithmic probability and the significance of the proofs on deep learning. It draws a parallel between the universality in Turing's proof on the universal Turing machine (UTM) and that in Solomonoff's proofs. It argues, in much the same way that the Church-Turing thesis gives an extreme outer limit of what it is possible to compute, Solomonoff's Algorithmic Probability gives an extreme outer limit on training a universal machine to discover regularities in a body of data. This chapter outlines the different philosophies and theories that contributed to his conceptualization of Algorithmic Probability. The outline helps us understand the core ideas of the abstract machine-learning model in Solomonoff's technical papers and how this model converges inductively with true probability. This convergence implies that one can always improve a machine-learning model by increasing the model's complexities and by training it with more data. This principle of convergence is essentially the basis for big data and deep learning. This chapter goes on to argue that the deep learning model satisfies the properties of the abstract machine-learning model in Solomonoff's Algorithmic Probability because the neural network in deep learning can compute any function like the UTM in Solomonoff's model. Acquiring an intuitive grasp of Solomonoff's proofs can help us verify the claim, often made by computer scientists, that AI empowered by deep learning can detect incredibly subtle patterns within large quantities of data.

Chapters 6 to 8 engage in a philosophical discussion on the social implication of cybernetics and AI. Chapter 6 presents an alternate perspective of cybernetics from Heidegger's dystopian critique of its universality. By alluding to Norbert Wiener's autobiography and to Gilbert Simondon's philosophies, the chapter argues that cybernetics is characterized not only by a reductionism of the living into mechanization and automation, but also by a philosophy of openness and complexity. This view aligns with our earlier discussion in Chapter 3 about the conflictual approaches held by early cyberneticians belonging to different scientific disciplines. In his endeavor to revolutionize scientific exploration, Wiener sought to theorize irregularities and complexities that escape the closed system of traditional scientific method. This attempt to theorize irregularities and complexities led to his conceptualization of cybernetics feedbacks and homeostasis. The chapter then introduces Simondon's theories of individuation and concretization while pointing out how these theories are associated with cybernetics. Simondon's philosophy is both an appropriation of cybernetics concepts such as negative feedbacks or information, and a critical response to cybernetics' mechanization of the living and its blurring of boundary between life and machine. The chapter draws from this critical response to discuss whether AI can be creative and exhibit life-like quality.

The openness of Simondon's philosophy of individuation is not purely technical. It theorizes how humans and technology can co-evolve by transcending conflicts and tensions. This open horizon and transcendence are only plausible due to the inherent potentiality in such tensions. Chapter 7 investigates further potentiality in Simondon's philosophy. Whereas Aristotle's model of potentiality was life, Simondon's model of potentiality was based on modern physical sciences. Quantum theories and solid-state physics explain observations that violate classical logic and the common sense of everyday life, and Simondon sees the necessity of a new philosophy that accompanies these seemingly non-sensical theories about the physical world. To attain a non-mathematical understanding of these twentieth-century scientific theories, this chapter draws from the explanations by Nobel laureate Richard Feynman, who is well-known for his knack of giving intuitive explanation behind the complex mathematical theories in physics. The understanding of quantum theories and solid-state physics can enlighten our interpretation of Simondon's concept of pre-individual reality. This enlightened interpretation would in turn reveal the significance of the figure-and-ground paradigm to Simondon's critique of technological alienation and why concretization is the means to overcome such alienation.

Chapter 8 compares and analyzes how Bernard Stiegler and Andrew Feenberg interpret and further develop Simondon's philosophy. The idea of concretizing the technical and the social has been raised by both Stiegler and Feenberg, but they have seemingly come to the opposite conclusions. Influenced by Heidegger's dystopian critique of cybernetics, Stiegler argues that a social world concretized into the global computational and information system would turn into the standing-reserves of a closed technical system. He points out that the channeling between social critiques and technical innovation is a necessary condition for the transductive operation of concretization and transindividuation, but this channel is being decimated by algorithmic governmentality. Feenberg, on the other hand, identifies the potentiality of transcending incumbent contradictions and stagnations in resistant social movements, in which the social and the technical undergo the transductive operation of concretization. This chapter argues, their seemingly opposite conclusion comes from the dual character of concretization as pathways to both openness and closedness, and the different emphases in Feenberg and in Stiegler reflect the former's focus on past social movements, which strengthened the channeling between social critiques and technical innovation, and the latter's forward-looking perspective of how algorithmic governmentality is decimating this channel. In this regard, the two thinkers hold a consistent interpretation of Simondon's philosophies in their respective social theories.

Leveraging the historical perspective and the philosophical arguments put forth in the preceding chapters, Chapter 9 examines whether advances in AI empowered by deep learning would necessarily bring about an automatic society that numbs individuals' capacity for social critiques, or if there can be alternatives for AI to co-evolve with the human and our society. The chapter addresses this question by identifying automation and symbiosis as two competing visions for computing research, and deliberates human-AI symbiosis as an alternative vision to AI automation. It argues that AI has no meaning outside the human context, and it is indeed possible that human societies may opt for an increasingly meaningless world under AI automation. But it is equally possible that societies may opt for an alternate path of sociotechnical development in which humans will evolve symbiotically with AI rather than becoming marginalized due to AI automation. To project how human-AI symbiosis may play out in real life, I draw from the science-fictions and the analysis of AI technologies in Kai-Fu Lee's and Qiufan Chen's *AI 2041* (2021). The short stories in *AI 2041* anticipate a future technological society built upon the principle of human-AI symbiosis and portray imaginary lifeworlds that allow readers to experience and to reflect on human-AI symbiosis. They exhibit how symbiotic relationship can exhibit a kind of creative openness, which seems to fulfill Simondon's vision of an open calculating machine with indeterminacy and programmability. Nevertheless, most stories in *AI 2041* took place in a society under surveillance capitalism or where AI technologies belonged to the social elites, illustrating that humans and AI can co-evolve symbiotically even when AI technologies remain alienated from the people. In this regard, the computer scientists' vision of human-AI symbiosis falls short of the revolutionary potentiality of Simondon's philosophy, which envisions not only an open machine but an open horizon of techno-cultural formation. His concept of pre-individual reality goes beyond the technical possibilities of human-computer symbiosis and posits that the human and the computing technology share a reality that is greater than whatever they appear to be. This implicates the always presence of potentiality for a historical unfolding of sociotechnical changes that transcend existing circumstances.

Chapter 1.

Methodology

1.1. Overview

This study conducts an inquiry into the technical lineage of machine learning, an investigation into formal proofs about its potentials, and a philosophical critique based on this immanent understanding. Toward the end of the 20th century, machine learning consolidated as a subfield of AI by drawing widely from algorithms across many fields of industry practice, including pattern recognition, signal processing, clustering, and computationally focused statistics (Wiggins & Jones, 2023, p. 181), and it has been the main driver behind recent advances in AI. Many are anticipating a proliferation of AI-related inventions that will have significant social consequences, both from a socioeconomic perspective (e.g., loss of jobs) and from the perspective of how human-machine interactions will be altered. This sentiment may be a consequence of the recent development in AI technologies, which include AI chatbots that manipulate political sentiment in the online world as well as the increasing ease of fine-tuning and customizing trained foundation models to generate models with domain-specific knowledge.[6] With such impending changes on the immediate horizon, traditional methodology in empirical research may be hard-pressed to keep pace with the social changes effected by new AI technologies. Conducting surveys and interviews helps reveal the locally situated perspective of the users. Such research is invaluable because the meaning of technology is inherently tied to the specific context of its deployment, but the efficacy of such research is limited to situations associated with technologies that have already been deployed. In anticipation of substantial social implications from AI inventions over the coming years, I want to conduct a critique of machine learning with a methodology that looks forward into the future.

My approach in this dissertation is to examine the basic ideas that have shaped AI in the past, as these same ideas should continue to shape AI in the future. It is

---

[6] For instance, see the Accelerate Foundation Models Research Program in Microsoft Research (https://www.microsoft.com/en-us/research/academic-program/accelerate-foundation-models-research-fall-2023/).

therefore crucial to immerse ourselves into the perspective of the early AI pioneers, the technical lineage that led to the birth of AI, and the rational basis behind their uncanny and bold claims. Critics of AI are prone to be befuddled by these claims without critically engaging with the formal proofs that lend support to these claims. The consequence is the conundrum of the two cultures[7], the culture of the intellectuals in humanities who imagine and criticize AI with free rein, and the technical culture cultivated in another silo, disregarding the voices of those outside the field as technically naive. To conduct an immanent critique, I want to become as knowledgeable as the AI practitioners about the fundamental ideas behind the functionalities of AI. Thus I investigate into the intellectual activities of the scientific community during the post-war period, since the field of AI is born out of a group of researchers grappling with a mixture of ideas from cybernetics and from universal computing during this period. This investigation ought to yield a technical understanding of AI through the lens of the experts in the field, which serves to either affirm or discredit controversial claims that have been communicated about AI.

After achieving an immanent understanding, I will step back to critically engage with the ideas that have shaped AI during this post-war period.[8] This critical engagement will be conducted philosophically. Between the traditions of analytic philosophy and continental philosophy, the former is closely associated with computational theories[9] whereas the latter brings a more critical perspective of technology. Hence continental philosophy appears to be more suitable for the purpose of a critical engagement with AI. Among the continental philosophers, this dissertation focuses on Martin Heidegger and Gilbert Simondon, a philosopher of Being and a philosopher of becoming. Heidegger's philosophy of Being led to his dystopian critique of a technocratic cybernetic society. He put forth the most influential critique of modern technology and was also the most vocal about the threats to humanity posed by cybernetics, which was instrumental to the development of AI and machine learning.[10] He famously claimed that cybernetics

---

[7] The idea reiterates the thesis of *The Two Cultures* (Snow, 1959/1993).

[8] The combination of an immanent understanding and an epistemic distance is somewhat similar to Pamela McCorduck's description of the type of person who can make the best predictions about AI: "[T]he best predictions of the distant future probably come from people who've been deeply immersed in the creation of such a technology, who are aware of its profoundest consequences, and who are then maybe slightly removed" (2004, p. 400).

[9] See Section 5.3 for the influences of Carnap's analytic philosophy on Solomonoff.

[10] See Chapter 3 and Chapter 4.

represents the culmination of the history of philosophy and that only a god can save us from the enframing of this technocratic society. According to some scholars, AI is the ultimate embodiment of Heidegger's dystopian critique of modern technology.[11] But as I will explain in Section 3.2, Heidegger's philosophical critique of cybernetics is based on a superficial understanding of cybernetics. In contrast, Simondon developed a philosophy of becoming by appropriating key ideas from cybernetics (see Chapter 6). His works exemplify how an immanent understanding of science and technology could serve as the basis of a critical philosophical inquiry of technology. In fact, they represent an important precedent for how I approach the immanent critique of machine learning.

The remainder of this chapter presents the methodology of this dissertation in greater detail. I begin with a literary review on the field of philosophy of technology, surveying the major works to provide the intellectual context for the philosophical discussion on Heidegger and Simondon. One of these philosophers, Hubert Dreyfus, was actively communicating with the AI community, but their mutual animosity is a telling story about the cultural gap widening between AI practitioners and humanities scholars.[12] I will provide an account on how this animosity developed and argue that the two sides could have come to a consensus if they could understand the reasoning of the opposite side. In my dissertation, I intend to move past this dichotomy, combining a critical philosophical inquiry of AI with an immanent understanding about the potentials of machine learning. I then explain why my project focuses on the intellectual activities during the thirty-year window of the post-war period, why continental philosophy is more suitable than analytic philosophy for conducting a critical inquiry of AI, and why the works of Heidegger and Simondon would be significant for a critique on AI. Toward the end of this chapter, I turn my attention from the past to the future and explain why my project will analyze science fictions that are written based on realistic projections of AI technologies.

---

[11] See Section 2.4.

[12] A majority of the early AI pioneers see Dreyfus as their nemesis. But as I will argue in Section 1.4, Dreyfus's critique genuinely exposes the limitation of symbolic AI but not AI in general. As the formalism of symbolic AI was the dominant approach at the time of his writing, Dreyfus was advocating a different direction for AI research. There are indeed researchers in the field of AI, such as Terry Winograd and Philip E. Agre, who were inspired by Dreyfus' critique of AI.

1.2. Literary Review of Philosophy of Technology

The immanent and philosophical critique of machine learning consists of the technological perspective and the philosophical perspective. The technological perspective focuses on the technical lineage and the computational theories of AI and machine learning. The philosophical perspective is situated within the tradition of philosophy of technology. This section provides the intellectual context to the philosophical discussion in this dissertation by surveying the significant works within this tradition.

The foundational text for philosophy of technology is Heidegger's essay "The Question Concerning Technology" (1953), as many later critiques of technology build on top of ideas from this essay. Heidegger contends that the sociotechnical world under modern technology has essentially become a closed system. He departs from the instrumental definition that sees technology as a means to an end, and argues that technology is "the realm of revealing, i.e., of truth" (1977a, p. 12). Whereas pre-modern technology reveals by "bringing-forth" a particular mode of living, in the sense of *poiesis*, modern technology reveals by "challenging" nature to the unreasonable demand of extraction and storage as technological resources (1977a, p. 14). This "challenging" mode of revealing becomes the horizon that both restrains and enables truth (1977a, p. 33). The harmony between pre-modern technology and nature is lost, and the world is now characterized by the teleological drive toward "enframing" all natural and human beings into "standing-reserves" (1977a, p. 17).

In this essay, Heidegger refers to the chalice to explain the causality in Greek thought and how modern technology deviates from it. We can trace the example of the chalice to Heidegger's discussion of craftwork in *Being and Time* (1927/2010). In this earlier text, Heidegger uses *Dasein*, meaning "there being" or "existence" in German, in lieu of the subject to express the unity between I and the world, and formulates the concepts of readiness-to-hand, presence-at-hand, and thrownness to distinguish the reality of everyday life from the domain of modern science. Readiness-to-hand describes the way *Dasein* unreflectively uses tools like a hammer, whereas presence-at-hand denotes how *Dasein* reflects upon the world as objectively present things. Ever since Descartes, the subject-object divide dominates the common-sense view of how the human mind and the actual world out there correspond to two isolated realities. Owing to this duality, objective analyses of the modern scientific discourse have become the dominant mode for making meanings and truths. Heidegger is neither satisfied with Descartes' dualism nor the unified theories of a purely subjective or objective reality (i.e., idealism or materialism) since Descartes. Rather, he addresses the subject-object antinomy by distinguishing beings that are ready-to-hand and those that are present-at-hand. He argues that seeing the world only in the mode of present-at-hand makes us blind to ready-to-hand relations, under which *Dasein* is "thrown" into a world of open possibilities that escape the formal present-at-hand analysis. In *The Question Concerning Technology*, Heidegger argues that people in the modern epoch are also "thrown" into a world where their understanding of the world is "enframed" as present-at-hand objects, as "standing-reserves," and their everyday life is "challenged" to conform to formal scientific "truths."

Heidegger's ontological critique of technology has been influential for prominent philosophers of technology in the latter half of 20th century. Among them are Don Ihde, Hubert Dreyfus, Albert Borgmann, and Herbert Marcuse. As Peter-Paul Verbeek (2006) explains, Don Ihde's analysis of human-technology relation in *Technology and the Lifeworld* (1990) is based on Heidegger's analysis of tools in the everyday relations between humans and their world. Ihde distinguishes three types of human-technology relations: the embodiment relation, the alterity relation, and the background relation. When people use technology in an embodiment relation or in a background relation, technology typically withdraws from people's attention. Hence Heidegger's readiness-to-hand, also characterized by the withdrawal of the tools, is the basis for Ihde's embodiment and background relation. In an alterity relation, technology emerges as the focus of attention, either when it breaks down or when it becomes a fetish of human desire. In either case, it does not facilitate a ready-to-hand relation between the user and the world. Technology appears as an objectively present being, and is therefore present-at-hand. This is how Ihde's "phenomenology of things" is influenced by Heidegger's distinction between readiness-to-hand and presence-at-hand.

Just like Ihde, Hubert Dreyfus in *What Computers Can't Do* (1972), which I will further discuss in Sections 1.3 and 1.4, derives his central argument from Heidegger's concepts of readiness-to-hand and presence-at-hand. Terry Winograd and Fernando Flores further develop Dreyfus's ideas in *Understanding Computers and Cognition* (1986). According to these authors, computers can only solve problems that are scoped within a clearly defined domain of possibilities, but they can never replace humans in encountering a lifeworld of contingencies and infinite possibilities. As such, terms such as “expert systems” and “decision support systems” are often misleading, as true experts in life are distinguished by intuitions that cannot be formalized. A doctor’s experience allows a shrewd discernment between a heart attack or chest pain due to stress, and a programmer guru can fix mysterious bugs in software development that nobody else can.[13]

Borgmann puts forth another Heideggerian argument in *Technology and the Character of Contemporary Life* (1988). He formulates the “device paradigm” that distinguishes between technological devices and what he calls “focal things.” Technological devices tend to reduce everyday life practices to a simplified form of a means to an end, as exemplified by microwave ovens as a means to cook fast food. The reduced form of living masks meanings in everyday life that are traditionally enriched by focal things such as a dining table. Such focal things set the stage for actions of care and concern, such as preparing for a meal and inviting friends over for dinner, that Borgmann calls “focal practices.” In the contemporary society, focal things and practices become increasingly marginalized, as the device paradigm is the dominant mode of living. Focal things and practices are derived from the concepts of readiness-to-hand, thrownness and care in Heidegger’s analysis. Similar to Heidegger, Borgmann sees the path toward liberation in the recovery of a “free” relation to technology. This “free” relation can be achieved by a way of living that prioritizes focal things and practices over the device paradigm.

In a Heideggerian critique, the path toward liberation typically involves changes in the way we live rather than in technology itself. This type of critique is often criticized as conservative and deterministic. In contrast, Marxism postulates how technology can be transformed into tools for liberation rather than for the reification of human lives. Herbert Marcuse is both a Marxist and a former student of Heidegger. According to Andrew Feenberg in *Heidegger and Marcuse* (2005), Marcuse’s *One-Dimensional Man*

[13] As I will explain in Section 1.4, the target of this phenomenological critique is symbolic AI. On machine learning and deep learning, a critique on the limitation of AI implemented by formalized rules is not applicable. See Sections 1.6 and 9.2 for discussions and examples on how AI empowered by deep learning is on the verge of replacing the functional roles played by doctors and software programmers.

(1964) indeed reiterates Heidegger's idea of a closed sociotechnical system under modern technology. But whereas Heidegger's critique is ontological, in the sense that the essence of technology is "enframing," Marcuse formulates a social critique. He identifies capitalism as the root of the closed form of rationality in the modern epoch. This apparent minor tweak to Heidegger's critique leads to an altogether different argument on how the human milieu may be liberated from the "enframing" of technology. The technological rationality that precludes values and meanings inherent in everyday life can be transformed into a liberated technological rationality with an aesthetic sensibility. This transformed aesthetic rationality can serve as the basis for a new science. As Feenberg points out (2005), Marcuse is vague on what constitutes this new science. His argument is much more appropriate for a transformed technology that encodes aesthetic values into its design, such as industrial machines that take environmental protection into account.

Marcuse is also a member of the Frankfurt School, and their Western Marxist tradition can be traced to Lukács' *History and Class Consciousness* (2013). Lukács sees the potential emancipation of the proletariat not only as a social phenomenon, but also as a philosophical resolution to the subject-object antinomy. It is in the praxis that the subjective reality and objective reality become one, and this unity is only achievable through the proletariat. Lukács contends that the "basic structure of reification can be found in all the social forms of modern capitalism (e.g. bureaucracy.)," but "this structure can only be made fully conscious in the work-situation of the proletarian" (2013, p. 172). The fetishism of the commodity reifies the lives of both the bourgeoisie and the proletariat, turning them into things that operate under abstract commodity relations in capitalism. But even though a worker is reified and dehumanized, "his humanity and his soul are not changed into commodities" (2013, p. 172). In contrast, a bureaucrat is reified even in thoughts and feelings. As Feenberg explains in *The Philosophy of Praxis* (2014), Marcuse generally agrees with Lukács' thesis on praxis as the resolution of antinomy, but for him, the site of praxis is not the revolution of the proletariat, but the revolution that transcends modern science and technology.

In Marcuse's attempt to formulate a new science and technology, he refers to *Du mode d'existence des objets techniques* (1958) by Gilbert Simondon. Just like Marcuse who refutes Heidegger's ontological claim, Simondon does not see technology as a threat in essence to the human culture. Instead, he identifies the source of technological alienation in the “opposition drawn between culture and technics, between man and machine” (Simondon 2016, 15–16). This man-machine opposition is akin to the subject-object antinomy that concerns Heidegger and Lukács. In parallel to how the concept of readiness-to-hand and the philosophy of praxis resolve the antinomy, Simondon formulates the theory of “concretization” to address the man-machine opposition.[14] Accordingly, technology does not evolve purely according to physical laws or in conformance to the subjective wills of designers or users, such as in slavery to the capitalist agenda. Rather, there is a certain nondeterministic logic that guides the co-evolution of man and machine. Technology evolves when elegant designs simplify its internal structure and its external interface with the milieu, which includes both the human and the natural milieu. This simplification makes the technology more concrete in the technical sense that it is more likely to survive than to be eliminated over time. The process of concretization is not purely subjective because elegant designs must take material constraints into consideration. But it is not deterministic either, as there is a manifold of possible paths that a technical evolution may take. Simondon theorizes the open co-evolution of man and machine in extension to his philosophy of individuation in *L’individuation à la lumière des notions de forme et d’information* (1964/2005). Building on Bergson’s *Creative Evolution* (1922), Simondon draws on the latest scientific discoveries in physics and biology of his time to formulate a material basis for the perpetual evolution of new forms. He identifies internal conflicts and contradictions as the sources of potentiality that empowers such open evolution. He then extends this theory of openness to work out a critique of modern psychology and social theory, and subsequently, to develop his philosophy of technology.

But Simondon was not the first to expound the concept of openness in technical design. This concept is also inherent in Karl Marx’s proposal of socialist technology in *Grundrisse* (1993). Marx argues that labour can become “attractive work, the individual's self-realization” under the right condition (1993, p. 611). This condition can be attained through a socialist transformation of the process of production. Workers under this transformed process will no longer be reduced to the status of a natural object. Rather, they can exert themselves as subjects who engage in “activities regulating all the forces of nature.” No longer reified, subjects can realize their potentiality through the serious

[14] I will deliberate on the philosophical significance of this concept of concretization in Chapter 8.

activities of labour. Hence Simondon's philosophy of openness is coherent with the theories of open technology in the Western Marxism lineage, which we can trace from Marx to Lukacs to Marcuse. These Marxists assert that technology does not necessarily determine the human milieu, but ought to evolve openly with it. They refute the technological determinism of Orthodox Marxism, which sees technology as the base that determines the superstructure of culture and social institutions. In comparison, social construction of technology (SCOT) also shares this refutation of technological determinism. They argue that the social meaning of any technology is not determined by its designers, but rather depends on the subjective interpretation of its users. Social constructivists such as Bijker (1987) conduct empirical and historical studies of technical artifacts such as the bicycle. These studies serve to bring us down to earth, from an abstract, macro discourse about technologies to the examination of the actors involved in shaping the design of particular technical artifacts. They attempt to show that there is no teleological essence to the evolution of technology, thus rejecting Heidegger's ontological critique of technology, and aim at illustrating the complex relations between actors behind each technical design.

The empirical research that employs the methodology of SCOT and ANT focus on disproving technological determinism and typically shy away from giving critical perspectives with much social significances. Ideology is a banned word in the SCOT and ANT research community. But the concept of absolute historicism in Marxism, in which truths are seen to be a product of history, is closely akin to social construction. It is in fact possible to combine the SCOT analysis and the Marxist critical theory, as Andrew Feenberg has done so in his various works (1995, 1999, 2002, 2017a). Following Marcuse, Feenberg argues that social values can be encoded in technology, which he calls "technical codes." The social adoption of a technology often comes from the implicit acceptance of these embedded technical codes as norms. When social actors participate in the negotiation of how a technology should develop, they bring their beliefs and values into such a process. Thus the negotiation is in fact a site of political contestation between different beliefs and values. The social construction of technology is the social determination of values and norms. In other words, technical systems, like social policies and institutions, are manifestation of power by the influential actors who shape technical development.

The above thinkers of technology are concerned with the dilemma that modern technology seems to both enhance the quality of living and suffocate the freedom of everyday life. Some of them identify the source of this dilemma in the subject-object antinomy. Others address the dilemma by distinguishing an open co-evolution of human and technology from technological determinism. Among these thinkers, Hubert Dreyfus was most active in his engagement with the computing and AI community. But their dialogues turned out to be far from cordial, and Dreyfus became AI's renowned archenemy due to their hostile and disrespectful interactions over the years.

1.3. Artificial Intelligence Needs a Good Dreyfus?

The discord of the two cultures between humanities and sciences is among the sharpest in the hostility between phenomenologist Hubert Dreyfus and AI practitioners. Dreyfus' paper *Alchemy and Artificial Intelligence* (1965) and his book *What Computers Can't Do* (1972) ridicule the grandiose predictions of AI researchers by comparing their wishful thinking to the misguided attempts of the alchemists. This alchemist metaphor surely antagonized the AI community, who in response scolded Dreyfus' ignorance about AI and computing. For instance, Edward Feigenbaum made the following comment about Dreyfus:

> What artificial intelligence needs is a good Dreyfus. … I can think of … one, maybe two philosophers who have the grasp of what AI and computing are all about, and also know philosophy. … But Dreyfus bludgeons us over the head with stuff he's misunderstood and is obsolete anyway—and every time you confront him with one more intelligent program, he says, "I never said a computer couldn't do that." And what does he offer us instead? Phenomenology! That ball of fluff! That cotton candy! (McCorduck, 2004, pp. 229–230)

Feigenbaum, along with other AI researchers, disapproved Dreyfus' critique on two accounts. The first reflects their disapproval of the phenomenological method. A "good Dreyfus" would presumably be an analytic philosopher who can further develop Carnap's digitalized representation of the universe or works on moral axiomatization of AI ethics. The second chastises Dreyfus' inability to "grasp what AI and computing are all about." When a critique of technology fails to recognize the potential capabilities of a technology due to its present limitations, the technical community would feel indignant and reject the critique as technically naïve. For them, engaging a meaningful dialogue with Dreyfus was impossible due to Dreyfus' premeditated stance against AI.[15] This view is epitomized by an anecdote recalled by Arthur Samuel (1974). Samuel was invited by Dreyfus to give a guest lecture at a philosophy class. He was surprised by the invitation and felt certain that Dreyfus did this with the sole purpose of humiliating him in front of the class: "[Dreyfus] thought he could shoot down all my arguments very well, and that he undoubtedly had his students primed to ask me questions" (Samuel, 1974). Samuel worked hard at anticipating every potential challenge that could be thrown at him. In his view, he was able to anticipate all the challenging questions and quickly gave concise answer to every question, taking pleasure at the sight of Dreyfus getting visibly upset.[16]

But as AI research endured what becomes known as the "AI Winter", it appeared that Dreyfus got the last laugh in the end. In *What Computers Still Can't Do* (1992), Dreyfus all but declared victory in his feud with the AI community:

> This edition of *What Computers Can't Do* marks not only a change of publisher and a slight change of title; it also marks a change of status. The book now offers not a controversial position in an ongoing debate but a view of a bygone period of history. For now that the twentieth century is

[15] In an interview by Pamela McCorduck, Samuel reflected on Dreyfus' philosophical differences with the AI community: "I think the trouble with [Dreyfus] is that he has certain philosophical beliefs and a person is entitled to your philosophical beliefs. We all have them, you have to have them in the real world. You have to have heuristics to live and some of the heuristics are these philosophical beliefs. What we're trying to do goes counter to his philosophical beliefs, and so, he tries to defend his philosophical beliefs by arguments, which I think are rather invalid. But he's so convinced that he's right, that he fails, or refuses to see his errors in his argument. Well, you just have to give up arguing with a person like that. It's like arguing about religion, you just don't argue about religion, because it's a matter of philosophical beliefs, and if a person has strong philosophical beliefs, he will only accept facts that fit his belief and he will always try to find devious explanation for apparent facts which contradict his beliefs. … I have a firm philosophical belief that a computer can do anything intellectual, in contrast with emotional, that a human mind can do, if we understand the process enough to write a program to do it. When people can solve a problem, it's an existence proof that it can be solved on a computer for me. That's my philosophical bent and if you present facts that invalidate that, I argue against them" (Samuel, 1974, pp. 34–35).

[16] Here is how Samuel told the story: "I've had a lot of talks with Hubert Dreyfus. I've had some very interesting experiences with him. He taught a course over at Berkeley and he invited me to come over and lecture to his course, which I was sort of surprised. I got to thinking about it, and I decided the reason he did that was, he thought he could shoot down all my arguments very well, and that he undoubtedly had his students primed to ask me questions. I really worked for his talk. I really worked hard thinking up all the things that I thought they would ask me and getting very glib, good and concise answers to these things and thought them through. They asked just almost every question, just down the line down the list and I had a pat answer and he was really upset about this and instead of getting me on the defensive, I got him on the defensive completely and I was really pleased about it. He has never invited me over since" (Samuel, 1974, pp. 33–34).

> drawing to a close, it is becoming clear that one of the great dreams of the century is ending too. Almost half a century ago computer pioneer Alan Turing suggested that a high-speed digital computer, programmed with rules and facts, might exhibit intelligent behavior. Thus was born the field later called artificial intelligence (AI). After fifty years of effort, however, it is now clear to all but a few diehards that this attempt to produce general intelligence has failed. (Dreyfus, 1992, p. ix)

Today, AI enthusiasts reading Dreyfus's premature "victory" statement would likely feel vindicated by the breakthrough of deep learning in AI research. For Dreyfus, the major roadblock in AI research used to be the lack of progress in pattern recognition: "It is not surprising, but all the more discouraging, that further progress in game playing, problem solving, and language translation awaits a breakthrough in pattern recognition research" (1965, p. 46). But it is the function of recognizing data patterns that deep learning excels at. Equipped with this pattern recognition capability, AI has become effective at language translation or in the game of Go, which is an ancient Chinese board game characterized by "virtually illimitable complexity" (Koch, 2016) and requires "true intelligence, wisdom, and Zen-like intellectual refinement" (K.-F. Lee & Chen, 2021, p. 7).[17] Such recent technical achievements attest to the verdict that Dreyfus had failed to grasp what AI or computing is all about. He had no appreciation of the computational theories that articulate what computing technology is potentially capable of doing, and Dreyfus was proven wrong time and again over his skepticism of what computers can do. Indeed, back in 1967, Dreyfus lost a chess match to the MacHack program two years after he remarked that "[s]till no chess program can play even amateur chess" (1965, p. 10). The AI community certainly took pleasure at this humiliating defeat. In the bulletin of the Special Interest Group in Artificial Intelligence of the Association for Computing Machinery, the results of the game were printed "with a headline drawn from 'Alchemy and Artificial Intelligence'—*A Ten-Year-Old Can Beat the Machine— Dreyfus*—and a subheadline that read, *But The Machine Can Beat Dreyfus*" (McCorduck, 2004, p. 232) Even though, as Dreyfus tried to clarify, his remark did not discount the ultimate possibility of a computer program playing reasonable chess, it is difficult to believe his clarification when, throughout "Alchemy and Artificial Intelligence," Dreyfus emphasizes

---

[17] As Kai-Fu Lee explains the significance of AlphaGo defeating the human Go champion in the Google DeepMind Challenge Match, "Go is a board game more complex than chess by one million trillion trillion trillion trillion times. Also, in contrast to chess, the game of Go is believed by its millions of enthusiastic fans to require true intelligence, wisdom, and Zen-like intellectual refinement" (K.-F. Lee & Chen, 2021, pp. 7–8). See also "How the Computer Beat the Go Master" (Koch, 2016).

the repeated failed attempts by AI researchers at developing chess programs and raises doubts that the method of symbolic AI would ever yield a computer program capable of playing a good match of chess.

To the early AI pioneers, Dreyfus' stubborn disdain against AI is akin to religious beliefs.[18] From their perspective, AI and alchemy cannot be more different. Whereas alchemy was grounded in a misguided mythology, the potentials of computing technology and AI are grounded in formal mathematical proofs. The logical foundation behind the limits and potentials of computing technology is the proof of universality of a Turing machine, and this proof was later extended by Ray Solomonoff to machine learning. The criticism of AI as a legacy of age-old human fantasy of anthropomorphism ignores the logical foundation of this "algorithmic" fantasy. Indeed, Dreyfus failed to recognize the potential of Solomonoff's mathematical theory of inductive inference, dismissing his theory as irrelevant on the ground of not having "a single example of actual progress," which to Dreyfus was a clear sign of stagnation in this line of research (Dreyfus, 1992, pp. 149–150). As I will explain in Chapter 5, Solomonoff's theory lends a theoretical credence to the vast potentials of machine learning. To be fair, at the time of Dreyfus' writing, Solomonoff's works remained in obscurity even among the AI researchers and did not garner the attention they deserve until the past couple of decades (see Section 1.5). Nonetheless, as this historical anecdote on Dreyfus demonstrates, a criticism on the computer's potential capability, based on what it can or cannot do today, likely cannot withstand the test of time.

1.4. Redeeming Dreyfus as the Prophetic Critic of AI

Dreyfus' bold claims and his agitating polemics certainly stirred up enmity and invited criticisms from the early AI pioneers, who were themselves guilty of enticing Dreyfus' animated hyperboles with their grandiose predictions about AI. In most people's eyes, "Dreyfus is persuaded that in the end artificial intelligence will never work" (McCorduck, 2004, p. 212). But a more careful reading of Dreyfus' critique would indicate that his critique was far from technically naïve as some of his critics had

[18] See footnote 15.

proclaimed.[19] Even though Dreyfus' critique on Solomonoff was based on the lack of evidential success in applying his theory of inductive inference, his phenomenological critique did expose the theoretical limitations of symbolic AI based on formalizable rules. In Dreyfus' central argument, humans use fringe consciousness in a pragmatic global context to recognize patterns in a way that both reduces and tolerates ambiguity. It is impossible to simulate this human ability with AI implemented by symbolic manipulation and formalizable rules, which was the dominant approach to AI research at the time of his writing. Accordingly, "the research program based on the assumption that human beings produce intelligence using facts and rules has reached a dead end" (Dreyfus, 1992, p. ix). So, the polemic of alchemy remains justifiable today because its only target of criticism was the method of symbolic AI, which is associated with "the belief that actions are governed by fixed values," "the notion that skills can be formalized," and "that one can have a theory of practical activity" (1992, p. 280). Dreyfus was contending that such a line of research would necessarily lead to stagnation, just like the fate of alchemy, and AI researchers ought to turn their attention to "fascinating new areas for basic research, notably the development and programming of machines capable of global and indeterminate forms of information processing" (1965, p. 84). "Indeterminate forms of information processing" is in fact the trademark of machine learning today! Rather than framing Dreyfus as the quintessential archenemy of AI research, he can easily be redeemed as the prophetic critic whose advocacy of "indeterminate forms of information processing" has finally been fulfilled by the recent innovations in machine learning and deep learning![20]

---

[19] Here is how Dreyfus recounts the type of criticisms he received on "Alchemy and Artificial Intelligence": "[T]he year following the publication of my first investigation of work in artificial intelligence, the RAND Corporation held a meeting of experts in computer science to discuss, among other topics, my report. Only an 'expurgated' transcript of this meeting has been released to the public, but even there the tone of paranoia which pervaded the discussion is present on almost every page. My report is called 'sinister,' 'dishonest,' 'hilariously funny,' and an "incredible misrepresentation of history. When, at one point, Dr. J. C. R. Licklider, then of IBM, tried to come to the defense of my conclusion that work should be done on man-machine cooperation, Seymour Papert of M.I.T. responded: 'I protest vehemently against crediting Dreyfus with any good. To state that you can associate yourself with one of his conclusions is unprincipled. Dreyfus' concept of coupling men with machines is based on thorough misunderstanding of the problems and has nothing in common with any good statement that might go by the same words'" (1992, pp. 86–87)

[20] His conclusion that AI research should be done on man-machine cooperation and symbiosis, which was defended by J. C. R. Licklider (Dreyfus, 1992, p. 87), is also a reasonable argument that I will further take up on later in this Chapter 9.

From this perspective, if the early AI pioneers could have toned down their overly optimistic predictions and if Dreyfus could be more cordial in his criticisms, it ought not be difficult for the two sides to arrive at mutual understanding and consensual agreement about the future of AI. Still, it is in fact a common practice that technologists routinely make exaggerated claims when they are under pressure to seek out research funding. If so, why would a philosopher like Dreyfus find the exaggerated claims about AI particularly disturbing? The answer seems to lie in certain characteristics unique to AI in comparison to other fields of technology research. As some critics have point out, there is a deceptive nature about AI from its very inception (Natale, 2021; Weizenbaum, 1976). There is a fine line between the computer science concept of abstraction, which translates between functional usage and implementation details, and the deceptiveness in AI, which elicits perceptive and emotive human responses that attribute non-existent capabilities to an AI program. This intrinsic deceptiveness was first noted by Joseph Weizenbaum on the effect his chatbot ELIZA had on his secretary who was testing the program. When Weizenbaum barged into the lab and interrupted the testing process, his secretary became visibly upset, as if Weizenbaum had disturbed her intimate conversation with a friend. "Chatting" with ELIZA aroused her emotional attachment even though she was well aware of the many hard-coded responses in this chatbot. From this incident, Weizenbaum realized that "extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people" (1976).[21]

Simone Natale (2021) pushes Weizenbaum's astute insight further, arguing that the power to induce delusional thinking has been central to AI's functioning throughout the history of its development:

> AI technologies entail forms of deception that are perhaps less evident and straightforward but deeply impact societies. We should regard deception not just as a possible way to employ AI but as a constitutive element of these technologies. Deception is as central to AI's functioning as the circuits, software, and data that make it run. … [S]ince the beginning of the computer age, researchers and developers have explored the ways users are led to believe that computers are intelligent. (Natale, 2021, p. 2)

[21] Douglas Hofstadter calls this the ELIZA effect (1995), and Hamid R. Ekbia calls this the attribution fallacy (2008).

Natale documented historical anecdotes to support his claim about the centrality of deceptiveness in AI research. These anecdotes demonstrate how AI "researchers and developers have explored the ways users are led to believe that computers are intelligent." Such deceptiveness indeed enticed early AI researchers to exaggerate the level of machine intelligence in their AI projects (see McCorduck, 2004, pp. 357–358).

At the same time, while it is fair to say that deception is "a constitutive element" of AI technologies," Natale probably goes too far in claiming that deception is central to AI. He argues that the Turing test reveals "the centrality of the human perspective" (2021, p. 31) and "the relationship between AI and deception" (2021, p. 32). For him, "[t]he playful deception of the Turing test, in this sense, further corroborates [his] claim that AI should be placed within the longer trajectory of deceitful media that incorporate banal deception into their functioning" (2021, p. 32). The early AI pioneers would likely have rejected this view of the Turing Test. They regarded Turing's seminal paper "Computing Machinery and Intelligence" (1950), which describes the Turing test, as the foundational pillar of the AI field. For them, the central question of "Computing Machinery and Intelligence" is, if a Turing machine can perform any describable function according to the proof of its universality, to what extent can it perform functions that humans would consider as intelligent? The paper explores the universality of Turing machine by conjecturing a paradigm shift on what intelligence is about and what kind of intelligence is a Turing machine capable of.

Rather than devoting their efforts to deceptive tricks, AI practitioners often see themselves as the creator of a new kind of intelligence, which may surpass human intelligence in certain aspects while lacking in other areas. Their ultimate goal is to create computing agents with general artificial intelligence, which is the ability to understand or learn any intellectual task that a human being can perform (Shevlin et al., 2019). In the eyes of AI enthusiasts, this aspiration is analogous to the quest for the holy grail. This quest is evident in Solomonoff's paper "The Time Scale of Artificial Intelligence: Reflections on Social Effects" (1985), which discusses six future milestones of AI. For instance, the fourth milestone is the ability to "read almost any English text and incorporate most of the material into its data base," which we seem to have achieved today with large language models such as OpenAI's generative pre-trained transformer (GPT). The fifth milestone is "a machine with a general problem-solving capacity near that of a human," and the final milestone is "a machine with a capacity many times that of the computer science community," which is later referred to as technological singularity. The critique of deceitful media only touches upon the tricks and hacks that make a program look human, but the AI community would consider these tricks as only peripheral activities complementing their core research, just as car accessories are considered peripheral functions in a car design.

How can we gauge whether an expert is giving a rationally sound projection based on actual potentials of the technology, or giving an unfounded speculation beyond its potential affordances? Are the last two milestones in "The Time Scale of Artificial Intelligence" more wishful thinking than expert projections with rational ground? Before rashly brushing them off as pure fantasy, we should note that the fourth milestone did not seem achievable at the turn of the century but is here with us today. In fact, throughout the history of computing, the sentiment on AI's potential capabilities has gone through cycles of optimism and pessimism, from the initial excitement on machines that exhibit aspects of human intelligence in the late 1970s and early 1980s (McCorduck, 2004, p. 417), to the stagnation in research and development during the so-called "AI winter" in the late 1980s (McCorduck, 2004, p. 418), then to the renewed optimism over the past decade due to the breakthrough in deep learning. According to AI expert Kai-Fu Lee[22], deep learning is the one and only breakthrough over the history of AI (K.-F. Lee & Chen, 2021, p. 536). Even though neural network is an old technology, it is only in recent years that deep learning, which uses massive amount of data to train neural networks with many layers, achieves its revolutionary status. As Lee describes, "[d]eep learning supercharged excitement in AI in 2016 when it powered AlphaGo's stunning victory over a human competitor in Go, Asia's most popular intellectual board game. After that headline-grabbing turn, deep learning became a prominent part of most commercial AI applications" (K.-F. Lee & Chen, 2021, p. 44). The supercharged excitement led to rejuvenated speculations about the potentials of AI. For some people, AI can one day attain human-level consciousness, while for others AI is no more than a collection of sophisticated tools serving human or social interests. From practical utilities to machines

[22] Lee is one of the most influential people in the field of Artificial Intelligence (Paz, 2020). He has been involved in AI research and product development at Apple, Microsoft, and Google, and managed $3 billion in technology investment as the CEO and the president of Sinovation Ventures (K.-F. Lee & Chen, 2021, p. 12).

with emotions and consciousness, there is a wide range of opinions on what AI can eventually become one day.

With such fuzziness in speculating about potential AI technologies, a succinct critique of AI ought to address this dilemma concerning the speculative potentials of AI. It is tempting to disregard all future prognosis about AI and directly examine the social implication of the technology in its immediate, current state. But speculative prognosis has a role in shaping the imagination of our future society as well as the actual designs of inventive technologies. There is in fact a collection of scholarly literature on the positive implication of speculative designs (Appadurai, 2013; de La Bellacasa, 2017; Dunne & Raby, 2013; Lupton & Watson, 2022). If we want to critically examine speculations about the future of AI, it would be best to avoid the same breakdown between Dreyfus and the AI community. In this regard, we want to stand in the shoes of AI experts and conduct an immanent critique. This does not mean that the predictions of AI experts should be taken for granted. What we can do instead, is to study AI by examining its technical lineage and by investigating the computational theories that establish the rational basis of AI's vast potentials. Our assumption is, the fundamental ideas that have shaped AI in the past will continue to shape AI in the future. Therefore, a philosophical critique of these ideas, uncovered in AI's technical lineage, will continue to be relevant in the foreseeable future of AI.[23]

1.5. An Immanent and Critical Inquiry

Attaining knowledge about a person's past can often help us appreciate the formation of her or his character. This is also true for the relation between a technology and its technical lineage. In this regard, researching the genealogy of machine learning is paramount to our objective of recognizing the ideas associated with machine learning. Over the functionalist history of this AI subfield, machine-learning techniques have evolved from the linear learning method with nice statistical properties before 1980s, to the non-linear decision trees and neural networks without the clean mathematical properties of linear methods during the 1980s, to the non-linear functions like support

[23] In a way, Dreyfus's approach is similar, but his argument was primarily an assessment about the potentials and limitations of symbolic AI, whereas we are assessing the potentials and limitations of machine learning, which Dreyfus briefly brushed aside due to its lack of pragmatic result in the time of his writing.

vector machines that have nice statistical properties in the 1990s.[24] But my project is not concerned with such a functionalist history, but with a technical lineage that traces the heritage of fundamental ideas in machine learning.[25] I want to explore the fundamentals of machine learning by examining what came before machine learning that shapes this field of technical research and how machine learning can be differentiated from this heritage. I also want to demarcate what machine learning can do based on formally proven computational theories. These intentions direct my attention to the vibrant scientific research activities in the thirty-year period during and after World War II, drawing together interrelated threads of intellectual movements between cybernetics, the early development of computing, and the birth of AI.

As a participant of the Macy Cybernetics Conference remarked, cybernetics "is a term that means all things to some men and nothing to many" (Grey-Walter, 1953/2003, p. 689). Critiques of cybernetics as an overarching social phenomenon under which all fields of knowledge will become branches of science include those of Heidegger (see Chapter 2), which views cybernetics as a universal science, and Bernard Dionysius Geoghegan's (2023), whose archival research suggests that the post-cybernetic technocratic reforms and scientific administration in France had huge implications for the rise of French social theory in the 1960s. Alternatively, we may view cybernetics more narrowly as a technical field, as a vague collection of technical concepts such as feedbacks or homeostasis, prior to examining the philosophical significance of such concepts. This is in fact the approach taken up by Simondon in his philosophical formulation (see Chapter 6). Following Simondon, my dissertation adopts this narrower,

---

[24] For an account of such a functionalist history of machine learning, see *How Data Happened* (Wiggins & Jones, 2023, Chapter 9), *The Master Algorithm* (Domingos, 2015), or *The Deep Learning Revolution* (Sejnowski, 2018).

[25] I adopted the term "technical lineage" from *On The Mode of Existence of Technical Objects* (Simondon, 1958/2016, pp. 44–48). In the section "Absolute origins of the technical lineage," Simondon illustrates what he means by technical lineage with the example of the diode, whose "*absolute beginning* [resides] in the association of this condition of irreversibility of the electrodes and of this phenomenon of transfer of electric charges through a vacuum: it is a technical essence that is created" (1958/2016, pp. 44–45), and "[a] technical essence can be recognized by the fact that it remains stable across the evolving lineage, and not only stable, but also productive of structures and functions through internal development and progressive saturation" (1958/2016, p. 46). Thus technical essence is analogous to the "crystalline germ" in Simondon's theory of individuation (see Sections 6.3 and 7.3 on the theory of individuation). My exploration of fundamental ideas behind machine learning is similar to Simondon's approach for identifying technical essence.

more technically-oriented perspective of cybernetics when examining the relationship between cybernetics and machine learning.

There are plenty of scholarly literature on the history of cybernetics (e.g., Dupuy, 2000; Geoghegan, 2023; Hayles, 1999; Mindell, 2002), computing (Aspray, 1990; e.g., Bardini, 2000; Bowden, 1953; Campbell-Kelly & Aspray, 2004; Turner, 2008), and AI (Cordeschi, 2011; e.g., Edwards, 1997; MacKenzie, 2004; McCorduck, 2004; Natale, 2021). These historical narratives provide holistic perspectives that identify institutions, people, and events behind the gradual emergence of these technical fields and shed light on the interwoven strands of ideas between them. For instance, David A. Mindell (2002) problematized the histories of computing that see the rise of computers first as logic machines before taking on cybernetic characteristics (2002, p. 10). He instead emphasizes the material substrate of the history of computing, reattaching computers to the problems and techniques, engineers and industries from which they sprang (2002, pp. 306, 317). He traces the history of control systems, including all kinds of feedbacks in circuitry designs that predate cybernetics. This history, which also documents the redeployment of telephony relays for the construction of digital relay computers, "reveals modern computing as part of a larger story of technology and culture, rather than the product of a discontinuous break between old and new" (2002, p. 317). In another historical account that sheds light on the genealogy of AI, Roberto Cordeschi (2011) explores the various historical stages "in the discovery of the artificial, both before and after the advent of cybernetics" (2011, p. 241). These stages include the study of living organisms as chemical machines, the origin of connectionism in the neurological hypothesis, and the association between behaviorist psychology and the robot approach.

The above accounts provide important historical context for my dissertation. Nonetheless, my focus in this dissertation is to trace the technical lineages of fundamental ideas that machine learning inherits from cybernetics. I am also less interested in the evolution of the modern computers and their hardware components, and more in abstract concepts and algorithms of machine learning, even though abstract programming is only possible due to hardware innovation that separates customization and programming from the mechanical limitations of hardware design (Mindell, 2002, p. 307). In addition, with the way cybernetics and computing overlapped and mutually influenced each other's development, historical accounts such as Mindell's typically place a greater emphasis on the complex history of interwoven relations between the two fields. Nonetheless, I will try to clarify their mutual influences by unraveling the interwoven strands of thought to elucidate the various threads of ideas associated with these unique technical fields in Chapter 3 and Chapter 4. The elucidation would be helpful for the later philosophical critique of sociotechnical potentiality associated with machine learning developed in Chapter 9.

With the birth of AI, pioneers of machine learning were exploring how a universal computer can be made to "learn." While Arthur Samuel was involved in developing software that exhibit this dimension of "learning," the burden of Ray Solomonoff was to come up with formal mathematical proofs on what machine learning is theoretically capable of "learning." Solomonoff's theoretical works did not garner the recognition they deserve in the AI community until the past couple of decades, as deep learning propelled machine learning to the forefront of AI research. But for the critical scholars in humanities, Solomonoff remains a relatively unknown figure, as the mathematical nature of his works makes them incomprehensible for non-mathematicians, which ironically include many AI practitioners working in the industry. The implications of Solomonoff's theory have been diffused into the AI community as common-sense knowledge. But some AI practitioners may exaggerate their claims about the potential capability of machine learning. Hence the pre-requisite of a critical inquiry of machine learning is to understand the potentials and limitations of machine learning based on Solomonoff's formal proofs. In Chapter 5, I will attempt to explain the reasoning of his algorithms and proofs in more-or-less everyday language. This explanation would contribute significantly to moving past the dichotomy between humanities and computer science, and to the bridging of computational theory and the rest of computer science. Scholars of humanities can collaborate with AI experts to formulate critiques on machine learning, without succumbing to the same antagonism between Dreyfus and early AI pioneers. The understanding would allow AI practitioners to understand the rationality and the boundary of their bold claims.

According to Solomonoff, the analytic philosophy of Rudolf Carnap had been very influential to his works (see Chapter 5). In general, analytic philosophy introduces two perspectives about AI. First, it fundamentally aligns with AI, as it contributes to the epistemology of AI. Carnap's writing on the logical foundations of probability and his view that the entire universe can be represented digitally have had direct implications on the epistemology and the algorithmic modelling of AI pioneers such as Solomonoff. In other words, the digital and probabilistic representation of the world in computing and AI is derived from Carnap's analytic philosophy. Second is an axiomatic study of ethics or moral philosophy, which is "the discipline concerned with what is morally good and bad and morally right and wrong. The term is also applied to any system or theory of moral values or principles" (*Ethics | Definition, History, Examples, Types, Philosophy, & Facts | Britannica*, 2023). For instance, "The Global Landscape of AI Ethics Guidelines" (Jobin et al., 2019) reviews 84 ethics guideline for AI and find 11 clusters of principles: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, solidarity. Luciano Floridi and Josh Cowls analyze forty-seven principles from a variety of sources (e.g., Future of Life Institute, 2017; Lords, 2018; *Partnership on AI*, n.d.; *The Declaration - Montreal Responsible AI*, 2017; Shahriari & Shahriari, 2017) and propose an ethical framework based on the six sets of principles for AI: beneficence, non-maleficence, autonomy and justice, and explicability.

In this dissertation, I want to explore beyond the realm of analytic philosophy to formulate my critique of AI, which draws on philosophy of technology, particularly the works of two continental philosophers in Heidegger and Simondon. Despite the dismissal of phenomenology as "cotton candy" by representatives of the AI community (see Section 1.3), continental philosophy, which includes phenomenology and critical theory, can contribute to a critical inquiry about AI. This branch of philosophy explores an alternate form of rationality beyond the scientific rationality that has become the dominant form of rationality over the past few centuries. In the context of computing and AI, the digital representation of the universe and the representation of our lifeworld in categorical fragments are not commeasurable with the phenomenological lifeworld in Heidegger's philosophy.[26] Continental philosophy typically questions the implications of this digital epistemology on humanity and society. It does not align with AI like Carnap's analytic philosophy. And unlike the principle-based AI ethicists, continental philosophers focus their attention on human-technology relations. They are typically interested in preserving the human agency in a world dominated by technology, theorizing concepts

[26] As Heidegger and Carnap are acquaintances whose philosophies are in disagreement (Dresser, 2020), digging deeper into their philosophical differences (Friedman, 1996, 2000; Nelson, 2013; Stone, 2017) may further enrich our critique of AI, but again such an effort is beyond the scope of this dissertation.

such as enframing, reification, and alienation.[27] Due to its epistemic distance with AI, continental philosophy can be critical of AI in ways that analytic philosophy cannot.

During the post-war period of technical advances, continental philosophers were alarmed by the positivism associated with such advances and formulated social critiques directed against the closed form of industrial technology and the iron cage of technocracy. Whether it be Heidegger's enframing of modern technology, Lukács' reification, or Marcuse's technological rationality, their critiques appear to be influenced by how humans were affected by the industrial technology such as the assembly line or by the systematization of bureaucracy. But as cybernetics and universal computing appeared to revolutionize technological development, these emerging technologies in the mid-twentieth century were afforded with the flexibility, adaptability, and contingencies of living organisms. This reframing of technology from inflexible to customizable or personalized machines has an ambivalent implication on how one may critique technology. One the one hand, with the blurring of the boundary between the living and the machine, humanity may become further entrenched in an overarching technological system. The appearance of flexibility and personalization in technology further co-opts human individuals to adopt and accept their ways of living within an overarching technological system. It is in this sense that, in Heidegger's philosophy, cybernetics represents the final stage of modern technology, as the fundamental science that brings the materialistic embodiment, and therefore the culmination and the end, of western metaphysics.[28] On the other hand, a person interacting with a cybernetic machine is less restrictive than interacting with industrial machinery, and the added

---

[27] This is similar to the view of Yuk Hui. For Hui, "[t]he fundamental question [in Heidegger's philosophy] is the regrounding of technology. We have to emphasize that this is not to add an ethics to AI or robotics, since we won't be able to change the technological tendency by just adding more values. Instead we have to provide new frameworks for future technological developments so that a new geopolitics can emerge …" (2019, sec. 44). He further argues, ""[W]e cannot ground morality on analytics, unless we believe in the kind of axiomatization we hear coming from the ethics of technology today. However, when we talk about the ethics of technology, we have already presupposed a specific kind of subject of knowledge and reasoning and assumed a certain normativity. Instead of axiomatizing the moral, we will have to go back to different modes of knowing which have yet to be taken into consideration by engineers and scholars working on artificial intelligence" (Hui, 2021, p. 353). His primary concern is human-technology relations and alienation: "[A] theoretical attempt to bring forward an ethics against anthropocentrism … will fail if it does not take the trajectory that we are going to outline—that is to say, the study of the human-machine relation—into account. … [Our project] seeks, at every opportunity, to allow spirit to exercise its freedom without producing the alienation of the soul" (2019, sec. 6).

[28] For more on this, see Section 2.2.

flexibilities facilitates a more open and therefore healthier human-technology relation. This positive take seems to align with the progressive aim of critical theory to transform science and technology so that people can be liberated from reification or the one-dimensionality of technological rationality. Simondon was the only continental philosopher scientifically and technologically savvy enough to deliberate this openness of cybernetics. He basically developed his entire philosophy by appropriating a number of cybernetics concepts around his theories on openness and potentiality. He reproached the dystopian critiques of technology contemporary to his time, like those by Heidegger.[29] Instead, he aimed at overcoming the antagonism between culture and technology. By drawing on his wide breath of knowledge on sciences (such as solid-state chemistry, quantum physics, biological evolution), and technologies (from mechanical engineering to electrical engineering to cybernetics), he brought together this technical perspective with a philosophy of individuation in order to explain how humans and technics co-evolve.

In fact, Simondon's writing exemplifies how one may conduct an immanent critique of technology. Simondon's critique is immanent in the way he often goes into scientific and technical details to identify their positive potentials and their negative aspects. As I will demonstrate in Chapter 6 and Chapter 7, his philosophical works ought to be read in parallel with the technological and scientific development of his days, which include cybernetics, quantum theory, and solid-state physics. These two chapters explore the implications of these fields for Simondon's philosophy at a level of detail beyond the existing commentaries of his works. At the same time, Simondon's writing brings a critical philosophical perspective of his immanent understanding of technology. His philosophical critique of Aristotle's hylemorphism, together with his immanent understanding of technology, serve as the basis for his conceptualization of his philosophy of individuation. In this regard, Simondon's immanent and philosophical critique of technology serves as a model for the critique of machine learning in this dissertation.

The difference between the Heideggerian critique and critical theory of technology resurfaces in the recent works of Bernard Stiegler and Andrew Feenberg.

[29] The more positive outlook for technology in Simondon is echoed in the critical theory and technical politics of Herbert Marcuse (1964, 2000) and Andrew Feenberg (1995, 1999, 2002, 2014, 2017b), and Langdon Winner (1978, 1993, 2010; Winner et al., 1997).

Both philosophers have grappled with the thoughts of Heidegger and Simondon to formulate their theories on contemporary technics, and their works share many common traits and converge in key ideas. These commonalities include the inseparability of the human and technics, the enframing of lifeworld by sociotechnical ensembles or technosystems, and the possibility of emancipation from within the technosystems rather than advocating a reversal to pre-modern technics.[30] Both philosophers have also further developed Simondon's project of overcoming the antagonism between culture and technics. Nevertheless, their philosophies set out from a different point of departure: Stiegler from an anthropological elaboration on Heidegger's thesis on cybernetics and Feenberg from an endeavor to continue Marcuse's incomplete project of left-wing technical politics. This difference is clearly revealed in their approaches to emancipation. Stiegler's "doubly epolkhal doubling" (Stiegler, 2016, p. 12) is an appropriation of Heidegger's salvation via an epochal change of collective awareness on the technological enframing of humanity. Human effort has limited contribution to this epochal change of collective awareness about the enframing essence of technology. In contrast, Feenberg's philosophy of praxis is realized in the technical politics of social movement. Human effort plays a primary role in bringing about the collective awareness and resistance against technological rationality. As I will further elaborate in Chapter 8, despite their similar interpretations about Heidegger's phenomenology and Simondon's theory of individuation, they emphasize different aspects on the theory and appear to draw opposite conclusions about the possible social implications and outlook of computing technologies.

To recap, clarifying the relationship between cybernetics and computing as well as understanding the theoretical potentials of machine learning can help us appropriate past philosophical reflections on cybernetics for our formulation of a critique on machine learning. Because machine learning has inherited ideas from cybernetics, we want to grapple with the philosophical significance of cybernetics through the writings of Heidegger and Simondon. Unlike analytic philosophers, these continental philosophers

---

[30] On the idea of emancipation from within technology, Feenberg uses the term "gestalt switch" (Feenberg, 2002, p. 16), which comes from Don Ihde's *Technology and the Lifeworld* (1990): "Any larger gestalt switch in sensibilities will have to occur from within technological cultures" (1990, p. 200). Stiegler develops a similar concept in "pharmacology" in *Automatic Society* (2016), which connotes the sense of drugs both for medication and for poisoning. So even though technology may appear to be poisoning society at the moment, the healing from such poisoning will also come in the form of technology as medication.

of technology theorize a critical perspective beyond scientific and technological rationality. And because Stiegler and Feenberg have further developed the philosophies of Heidegger and Simondon in their deliberations of the dystopian dimension as well as the liberative potentiality of modern technology, we will also refer to their works to develop a philosophical framework for a social critique of AI and machine learning.

1.6. Situated knowledge from an Imaginary Future

Turning our attention from the past and the present to the future, the dichotomy of the two cultures between the AI community and the intellectuals in the humanities, which we discussed in Sections 1.3 and 1.4, becomes even more pronounced over the wide spectrum of speculations about the future of AI. This dichotomy is clearly manifested in the confusion between rational projection based on scientific facts and futuristic imagination as a mixed bag of scientific-based interpolations coupled with pure speculations. Looking from the outside, technology critics in the humanities tend to be skeptical of the bold proclamations made by AI experts about the future of AI. But for those inside the circle of AI practitioners, they see themselves as someone in a privileged position, witnessing first-hand the latest advances in AI research that make certain proclamations seem justifiable. My approach is to take the claims and the visions of AI experts seriously but also critically, adjudicating and evaluating some of these proclamations by drawing on the aforementioned immanent critique of AI and machine learning, which includes a historical inquiry into the cybernetics movement and the early history of AI, an investigation into computational theories on the potentials of machine learning, and a critique of human-technology relation based on continental philosophy.

At the same time, if the scholars in technology studies are correct about the co-production of the social and the technical, about how the meaning of technology arises and evolves in relation to the specific cultural context of its deployment, then my project needs to take into account the situated knowledge that arises over the feedback reaction between technology and localized culture. There have certainly been ample empirical works on surveillance capitalism, algorithmic governance, autonomous driving, or other types of technosystem empowered by machine learning. Indeed, by studying the way machine learning shapes people's lives today, we can identify the immediate problems, such as discrimination, that are caused by the actual implementations of machine learning. But such identification only serves the formulation of a negative critique but cannot offer suggestions of changes from within the technological culture, of changes in technical design rather than in policies, of how a gestalt switch[31] may come about. Hence in my final analysis, I want to go beyond the predictions by AI experts on the future of AI technology. My plan is to examine science fictions on AI, which would presumably allow readers to immerse into a lifeworld to attain a holistic perspective of a future society populated with impending AI technologies.

Nevertheless, not all science fictions are written with insiders' knowledge about where the future of AI is heading. Science fiction about conscious AI androids with superhuman quality, like those in the *Westworld* TV series, would not meet the criteria of imagining a future world based on realistic AI. One book that meets such criteria is Kai-Fu Lee's and Qiufan Chen's *AI 2041* (2021). This book combines storytelling with technical insights to bring to life a world proliferated with futuristic applications of AI. It is composed of a number of short stories written by Chen, each story followed by Lee's analysis of the visionary technologies that shape the story as well as his anticipation of the technology roadmaps over the coming years. Riding on Lee's industry exposure, the authors distinguish their writing from other science fictions by claiming that the AI technologies anticipated in the book are based on prototypes already working in research labs around the world. They "avoid making speculative predictions about fundamental breakthroughs and rely mostly on applying and extrapolating the future of existing technologies," with the conviction that "even with few or no breakthroughs, AI is still poised to make a profound impact on our society" (2021, p. 12). Presumably, *AI 2041* helps readers "imagine the future of the world and our society in twenty years' time … to tell the 'real' AI story … This book is based on realistic AI" (2021, p. 11).[32]

[31] See footnote 30 for the definition of gestalt switch.

[32] According to Lee and Chen, people often exaggerate predictions that "miss the complete picture" (2021, p. 10) because they typically rely on three sources to learn about AI: science fiction, news and influential people (2021, p. 10). Science fiction books and TV shows often depict super intelligent androids turned evil; media reports tend to neglect incremental advances and focus on the negatives, such as misinformation and deep fakes or autonomous vehicles killing pedestrians; the predictions of "thought leaders," who are not experts in AI technology, often "lack scientific rigor" (2021, p. 10). In contrast to these sources, *AI 2041* is a project that attempts to "balance [social] concerns with exposure to the full picture and potential of this crucially important technology" (2021, p. 10). As Kai-Fu Lee explains, "Believers in singularity argue that exponentially improving technologies will lead to superintelligence. I agree that AI computational prowess has indeed increased exponentially, but exponentially faster computing power alone does not lead to qualitatively better AI. To deliver qualitatively better AI, new scientific breakthroughs like deep learning are also needed. Suppose we had all the computing power

Given that my exploration concerns the future of machine learning, the science fictions of *AI 2041*, presumably based on a realistic projection of AI technologies, become the site for my deliberation on the situated knowledge of future localized scenarios. These are the case studies that provide a realistic anticipation of the near future of AI technologies coupled with an imagination of how such technologies interact with our lifeworld. These case studies can desublimate our abstract discussions into a deliberation of life experiences. To get a sense of the imaginary lifeworld portrayed in *AI 2041*, I provide a survey of selected stories here, summarizing their storylines as well as the AI technologies associated with them. Toward the end of this dissertation, I will draw from and critically examine these stories, using them as the lifeworld context for the formulation of my philosophical critique on machine learning.

The story "The Golden Elephant" is about a deep-learning enabled insurance program called the Golden Elephant, which gives monetary incentives to those who allow the program to tap into their personal data stream. The program would make different recommendations about how to live well, and would penalize the insured participants if they engage in activities deemed as threatening to their well-being. For instance, an overweight boy eating more sweets would endanger his health, and an immediate hike would be reflected in the family's insurance premium for failing to stop the boy from eating. The kick of the story is how the Golden Elephant attempted to dissuade a girl from falling in love with a boy who belongs to a different caste, and how the young couple managed to break away from the confinement of the presumably intelligent insurance technosystem.

"Twin Sparrows" explores the potential advances in human-computer symbiosis with the availability of virtual companions. The AI companions in the story can converse fluently in human language, and can also act as smart AI teachers "camouflaged as virtual cartoonlike friends" (2021, p. 98). These cartoon like figures would appear on a person's glasses in an XR (Extended Reality) overlay. In the story, a pair of twins had

today but no deep learning; then the whole AI industry would be non-existent. To achieve superintelligence in the future, we absolutely need more scientific breakthroughs" (K.-F. Lee & Chen, 2021, p. 536). Even though we cannot discard the possibility of future scientific breakthroughs, it is still pointless to formulate critiques on wild speculations that may never happen or are, as Noam Chomsky remarks, eons away from happening (Socrates, 2013). Nevertheless, it is possible to anticipate the impending changes based on the existing state of technologies without major breakthroughs, while assuming the highly plausible increase in computational power.

lost touch with one another, but they eventually yearned for a re-encounter with each other due to their telepathic-like visions. As the story later reveals, these visions via the XR glasses were not truly telepathic, but were generated by their AI buddies, whose underlying code had been secretly embedded with a communication protocol between them. This protocol was programmed by an IT staff of the orphanage that took care of the twins during their childhood.

"Contactless Love" explores how COVID-19 has accelerated trends in new drug discovery, precision medicine, and robotic surgery, all enhanced by AI (2021, p. 164). One of "AI's greatest achievement for science to date" is its ability to determine protein folding, which is an essential step in drug discovery (2021, p. 206). Wearable devices can help AI correlate the statistics from monitoring heart rate, blood pressure, and other vital statistics for early detection and precise medical treatment. Rather than training AI from expert medical knowledge, AI is now being "trained directly from real patient-treatment-and-outcome data" (2021, p. 203). Lee also anticipates that fully autonomous robotic surgeries will increasingly account for the majority of procedures (2021, pp. 206–207), and "diagnostic AI will exceed all but the best doctors in twenty years," with this trend first being felt in radiology, pathology and diagnostic ophthalmology (2021, p. 208). Eventually, human doctors will be "transformed into something akin to compassionate caregivers and medical communicators" (2021, p. 206).

"The Holy Driver" imagines a society in the midst of transitioning from human drivers to autonomous driving by AI. Due to AI's inherent limitations, young talented gamers were recruited for mysterious projects of game playing. As it turns out, the gamers were actually saving the lives of real people from natural disasters or acts of terrorism. Their car racing games, immersed in virtual reality, were live replicates of situations somewhere in the real world, with actual cars being remotely controlled by these gamers in real-time. As Lee explains, even when autonomous driving is mature to the point where no human is required for any roads and environments, and become safer than vehicles with human drivers in standard situations, there are still problems that "could potentially confused the AI, including natural disasters or acts of terrorism" (2021, p. 316). Human-level dexterity is required in these scenarios, and "the best solution would be to bring an expert human into the car and take over" (2021, p. 316).

The story “Quantum Genocide” gives us a glimpse on the potential impact of quantum computing and AI-enabled autonomous weapons. The villain of the story is a computer scientist who turned mad after his wife and daughter died from a wildfire. He blamed the entire human race for this climate-change related personal tragedy, and embarked on a revenge plot against all of humanity. The story illustrates how quantum computing “could turbocharge AI and computing” (2021, p. 324), and how the race for quantum supremacy can have huge political significance in the not-so-distant future. Computational power will be the determining factor on who wins the cyberwarfare and the technical arm-race, in which swarms of AI drones will become the most lethal assassins. The story also raises our awareness on the existential crisis associated with autonomous weapons whose prowess “largely comes from the speed and precision from not having a human in the loop” (2021, p. 390).

“The Job Savior” explores what will happen to human jobs when AI seeps into more industries and makes human tasks redundant. It describes the emergence of a new industry, job reallocation firms, which are called on to retrain and reassign displaced workers. The job reallocation firms often need to repeatedly face the same workers finding themselves out of jobs again and again. Due to new advancements in AI, their reassigned jobs would be supplanted by AI after only a handful of years. This story reveals the authors’ belief that “artificial intelligence can perform many tasks better than people can, at essentially zero cost” (2021, p. 429). Due to AI’s edge over humans “in its ability to detect incredibly subtle patterns within large quantities of data” (2021, p. 430), the authors anticipate that “AI will be doing everything from underwriting our loans to building our homes, and even hiring and firing us” (2021, p. 429).

In all these stories, Lee and Chen confine their imagination to their realistic prognosis of AI-related inventions. As Lee explains, there are three human capabilities where AI falls short of today and will likely struggle to master even in twenty years. First is creativity: “AI cannot create, conceptualize, or plan strategically. While AI is great at optimizing for a narrow objective, it is unable to choose its own goals or to think creatively” (2021, p. 435). Its second limitation is empathy: “AI cannot feel or interact with feelings like empathy and compassion … [it] cannot make another person feel understood and cared for” (2021, p. 435). Third is dexterity: “AI can’t deal with unknown and unstructured spaces, especially ones that it hasn’t observed” (2021, p. 435), as is the case with autonomous driving. In contrast to techno-posthumanism and technological singularity, Lee and Chan envision that "humans will work symbiotically with AI, with AI performing quantitative analysis, optimization, and routine work, while we humans contribute our creativity, critical thinking, and passion" (2021, p. 11). This symbiotic relationship is most evident in the "Twin Sparrows," where AI companions are consigned meanings by their human buddies; they are not standalone software agents with consciousness. In this sense, the stories in *AI 2041* are distinct from popular science fictions such as the movies *Her*, *Ex Machina*, or the *Westworld* TV series, in which highly intelligent androids are portrayed as creative, affectionate, and capable of attaining self-consciousness. We will come back to analyze this symbiotic partnership between AI and human's creativity, dexterity, and empathy toward the end of this dissertation in Chapter 9.

1.7. Summary

This chapter presents my overall approach in formulating an immanent critique of AI. It begins with a survey of the field of philosophy of technology. The hostility between one of these philosophers, Hubert Dreyfus, and the early AI pioneers marks the beginning of the cultural gap between humanities and the AI community. This hostility attests to the importance of attaining an immanent understanding of machine learning. While Dreyfus was astute in pointing out the limitations of the early formalist approach of symbolic AI, his polemics and seemingly hostile stance against AI camouflage his prudent recommendation for AI research that addresses contextual regularity not governed by formalizable rules.[33] Moreover, he overlooked the significance of Solomonoff's theoretical works and thus failed to grasp the full potential of machine learning in addressing context regularity and pattern recognition.

While a critique that fails to grasp the true potentials of AI would easily be rejected as technically naïve by AI experts, these experts often make bold claims about the potentials of AI based on their exposure to state-of-the-art AI research rather than on rigorous arguments that substantiate their claims. The technical community may spread among themselves the implication of important computational theories, such as the

[33] "We shall now try to show not only that human behavior can be regular without being governed by formalizable rules, but, further, that it has to be, because a total system of rules whose application to all possible eventualities is determined in advance makes no sense" (Dreyfus, 1992, p. 257).

formal proofs of Solomonoff, without understanding or referring to the proofs. To attain an immanent understanding, my project involves a reading of Solomonoff's computational theories on machine learning. In addition, because ideas from the cybernetics movement and from universal computing have been instrumental to the birth and development of AI, unraveling the interwoven strands of thought and the mutual influences between cybernetics, computing, AI, and machine learning can help elucidate the various threads of ideas associated with these fields. Such ideas, which have fundamentally shaped AI in the past, should continue to shape AI in the future.

This immanent understanding ought to be accompanied by a critical perspective that stands at a certain epistemic distance from AI. Since AI adopts its digital epistemology from Carnap's analytic philosophy, continental philosophy would appear to be more appropriate for a critical inquiry of AI than analytic philosophy. The AI ethics in analytic philosophy is concerned with a principle-based ethical framework for AI, but this framework tends to evaluate AI categorically from within rather than holistically from a critical distance. In comparison, continental philosophy is chiefly concerned with a critique of human-technology relation at a sophistication beyond the category of autonomy. During the post-war period, philosophers from both phenomenology and critical theory were targeting their critiques on modern industrial technology. The assembly line or bureaucratic systematization were turning human subjects into resources or standing reserves in supportive of the overarching technological system. But as cybernetics and universal computing emerged during the post-war period, technology is reframed from inflexible to customizable or personalized machines. This reframing has an ambivalent implication on technology. On the one hand, critics such as Heidegger are alarmed by humanity becoming further entrenched in an overarching technological system. On the other hand, the new-found flexibility in technology appears to facilitate a more open and therefore healthier human-technology relation. This positive take seems to align with the progressive aim of critical theory to transform science and technology so that people can be liberated from the domination of an overarching sociotechnical system. This is why critical theorist like Marcuse or Feenberg found appeals in Simondon's works. Unlike most philosophers, Simondon was very knowledgeable in the latest development of science and technology at his time. He developed his philosophy of openness by appropriating cybernetics concepts and his model of potentiality from modern physics such as quantum theory. By thoroughly engaging with cybernetics and other branches of sciences, Simondon's works also exemplify how one may conduct an immanent critique of science and technology. The ontological critique of Heidegger and the progressive critique of critical theory represent the different points of departure for the philosophy of Stiegler and that of Feenberg. Despite sharing common interpretations of Heidegger and Simondon, they came away with different approaches to bringing emancipation from technological domination. Like Heidegger, Stiegler does not see human efforts as sufficient in bringing about an epochal change in collective awareness about the essence of technology. For Feenberg, human efforts via technical politics play a significant role in bringing about the change in collective awareness on the technological domination of humans.

Turning our attention from the past and the present to the future, the dichotomy of the two cultures between science and humanities becomes even more pronounced when it comes to speculations about the future of AI. While AI practitioners may exaggerate their claims or implement deception as inherent characteristics of AI, technology critics outside the field of AI may not be knowledgeable enough about the latest advances in AI, which often take place in research labs a few years before they are released to the public (as is the case of large language models over the past few years). My approach is to take the claims and the visions of AI experts seriously but also critically, adjudicating and evaluating some of these claims by drawing on the previously established immanent and philosophical critique of AI and machine learning. But rather than addressing literature that directly predicts the future of AI, I elect to analyze science fiction written based on realistic projections of AI. This would allow my project to account for the situated knowledge that arises over feedbacks between localized culture and technologies yet to exist.

As an addendum, with all the anxious reaction about the possibility of students' papers being written by ChatGPT, I decide to experiment with a somewhat playful ploy for the rest of the chapters in this dissertation. I will begin each chapter with a question-and-answer dialogue with ChatGPT. This scheme has the dual purpose of introducing a key topic or concept related to that chapter, and to confirm that the chapter's content has more substance than what this AI chatbot can generate. The ironic twist is the implicit critique that ChatGPT, in spite of the virality of its popular reception and reports of its multi-facet skills such as passing the Google Coding Interview for entry-level software engineers (Bhaimiya, 2023), cannot yet generate a PhD dissertation that makes original academic contributions.

Chapter 2.

Cybernetics and the Technological Will to Will

Me: what does Nietzsche mean by will to power?

ChatGPT: The "will to power" is a central concept in the philosophy of Friedrich Nietzsche, who describes it as the basic drive behind all human behavior, thought, and action. It refers to the instinctual urge to increase one's own power and control over one's environment, both internally through self-mastery, and externally through domination over others. The "will to power" is not just a desire for power in a conventional sense, but a deep-seated striving for excellence and self-overcoming. It is a fundamental aspect of the human condition and, according to Nietzsche, an essential part of life and growth.

Me: what does heidegger mean by "technological will to will"?

ChatGPT: The "technological will to will" is a phrase used by Martin Heidegger to describe the way in which modern technology shapes and influences our understanding of the world and our relationship to it. Heidegger saw technology not simply as a tool for achieving practical goals, but as a way of revealing and shaping the world that is essentially bound up with our own will to power.

According to Heidegger, the technological will to will is characterized by a drive to control and dominate the world, to turn everything into a resource for our own use, and to see everything in terms of efficiency and productivity. He argues that this way of thinking and being is a direct result of the modern understanding of technology as a means to an end, rather than as an end in itself. Heidegger argues that this view of technology leads to a fragmented and alienation view of the world, where we no longer see the interconnectedness of things, but only see them as means to our own ends.

2.1. Overview

In an interview given by Martin Heidegger to Rudolf Augstein and Georg Wolff for *Der Spiegel* magazine (Alter & Caputo, 1976) on September 23, 1966, Heidegger lamented that “everything is functioning and that the functioning drives us more and more to even further functioning, and that technology tears men loose from the earth and uproots them" (1976, p. 277). This world movement brings about "an absolutely technological state" to the point that "[o]nly a god can save us" (1976, p. 277). He was then asked "what or who takes the place of philosophy" as that which shapes the horizon for effecting changes in the western world, and he answered with a single word: "Cybernetics" (1976, p. 279). This answer highlights the significance of cybernetics for the late Heidegger, not just as a fundamental science but as the impetus behind a technocratic world movement that enframes humanity in a technical state. As Yuk Hui remarks, Heidegger claims that the emergence of cybernetics in the mid-twentieth century marked the completion and end of philosophy (Lovink, 2019). Heidegger presents his argument on cybernetics in a lecture given in 1964, entitled "The End of Philosophy and the Task of Thinking" (2002). In the lecture, he contends that philosophy as Western Metaphysics, which "thinks beings as being in the manner of representational thinking which gives reasons" (2002, p. 56), is coming to completion in the development of the empirical sciences of man "determined and guided by the new fundamental sciences which is called cybernetics" (2002, p. 58). This view on the completion and end of metaphysics in cybernetics is already implicit in "Overcoming Metaphysics" (Heidegger, 1973b), a collection of notes from the years 1936 and 1946. The notes contain his reflections on the historical development of the will of representational thinking, culminating in the extreme form of Nietzsche's will to power as the technological will to will. This "will to will forces the calculation and arrangement of everything for itself as the basic forms of appearance, only, however, for the unconditionally protractible guarantee of itself" (Heidegger, 1973b, p. 93).

Through wrestling with the will of representational thinking in western metaphysics, Heidegger concludes that cybernetics supplants western philosophy to the point that only a god can save us. Due to this much-maligned remark about a god, many critics are more than ready to discard Heidegger's critique of technology because human agency seems to play no role in shaping technology under his overzealous dystopian determinism. But his brief responses in the interview are in fact only the simplified conclusions of his elaborated philosophical expositions from his later writings. Therefore, to fully understand Heidegger's responses, it is imperative to interpret them in light of his extended philosophical works. We ought to break down the abstractness of his claims and immerse deeply into his thoughts. Doing so can help clarify what Heidegger means by the phrase "only a god can save us," in what sense would cybernetics take the place of philosophy, and what he means by the "technological will to will."

In this chapter, we will address these questions by reading "The End of Philosophy and the Task of Thinking" (Heidegger, 2002) and by navigating Heidegger's critical reflections on Nietzsche's will to power and on the technological will to will. There are ample commentaries on Heidegger's interpretations of Nietzsche. Among them, this chapter draws on Bret W. Davis's *Heidegger and the Will: On the Way to Gelassenheit* (2007), which I find most comprehensive and illuminating for the purpose of understanding Heidegger's argument on the history of metaphysics as the history of the will. Davis shows us the development of Heidegger's thoughts across his life time, marked by the "turn" during his lectures and writings on Nietzsche. Heidegger went from the articulation of a "proper will" to a "non-willing" as a fundamental (dis)attunement of letting beings be, of releasement-toward-things (*Gelassenheit zu den Dingen*), which serves as the alternative to the willfully positing of beings in cybernetic calculation and arrangement. For Heidegger, the extreme epoch at the end of the first beginning of the history of Western metaphysics is ironically the tipping point for an other beginning of the history of being marked by the *Gelassenheit* of "letting beings be." This eschatology of the transition from the epoch of the will cannot be "willed" by human beings, but we may nonetheless participate in this transition between epochs.

Heidegger's reflections on cybernetics and the technological will to will has been taken up by Michael E. Zimmerman (2016), who interpolates what Heidegger might have said on the techno-posthumanism that Ray Kurzweil proposed in *The Singularity is Near* (2005). For Zimmerman, "Heidegger would say that techno-posthumanism is the latest and perhaps most dangerous phase in the era of techno-industrial nihilism" (2016, p. 101), and "[s]uper AI would be, in effect, the ultimate ontical embodiment of what Heidegger—drawing on Nietzsche—calls the Will to Will" (2016, p. 101). This notion of techno-posthumanism seems prevalent in the writing of Stiegler and Hui, whose works repeatedly bring up the posthumanist ideology in a Heideggerian sense. Hence this reading of Heidegger would serve our dialogues with Stiegler in later chapters.

In the following, I will first elaborate on why cybernetics marks the completion and the end of philosophy for Heidegger. I will then go into Heidegger's interpretation of Nietzsche's will to power, the history of metaphysics as the history of the will, and the technological will to will. This paves way for a discussion on how Heidegger's critique on technology can illuminate us on techno-posthumanism. I will conclude by reflecting on Heidegger's *Gelassenheit* and his eschatology of an other beginning.

2.2. Cybernetics as the End of Philosophy

In "The End of Philosophy and the Task of Thinking" (2002), Heidegger wrote, "[p]hilosophy is metaphysics. Metaphysics thinks being as a whole—the world, man, God—with respect to Being, with respect to the belonging together of beings in Being" (2002, p. 56). Etymologically, the prefix "meta-" in metaphysics means "beyond" or "after." Metaphysics presumes the existence of some reality beyond the physical reality and contemplates about this "meta-physical" reality that cannot be observed with our senses. This presumption originates in Plato, as is evident in the allegory of the Plato's cave, and western philosophy since Plato has been an ongoing dialogue about this metaphysical reality. Rather than empirically examining "what is present," as in physics, metaphysics contemplates "what is present" in terms of the reality beyond our senses.

In Heidegger, "what is present" is referred to as "beings," and the reality beyond our senses is contemplated in his thoughts on the "Being of beings." What does Heidegger mean by the "Being of beings?" We can find his clarifications in "Metaphysics as History of Being" (1973a):

> "Being" means that beings are, and are not nonexistent. "Being" names this "That" as the decisiveness of the insurrection against nothingness. Such decisiveness emanating from Being at first arrives in beings, and here adequately, too. In these beings Being appears. So decisively has Being allotted beings to itself (in Being) that this does not need to be thought expressly. Beings give adequate information about Being. (1973a, p. 1)

"Being" is the overcoming of nothingness in the presencing of beings. It is the horizon for beings to be revealed. In his writing, Heidegger adopts the terms "beings" and "Being" from early Greek philosophers. He uses "beings" instead of "subjects" or "objects" to recover our thoughts on "what is present" prior to the subject-object split of beings since René Descartes (1973c, p. 69). He elects the term "Being"[34] over "the God of onto-theology" in order not to "remain bound to a history of metaphysics as onto-theology

[34] For philologists, "Being" in Greek implies deity.

(Davis, 2007, p. 124). As Bret W. Davis explains, “Heidegger’s god(s) is no more the almighty Creator of heaven and earth than it (they) is an eternal being of self-presence” (2007, p. 249). Thus “insurrection against nothingness” does not result from the God of creation, but rather emanates from Being, which is an abstract philosophical construct representing the ground from which beings become present. As the ground of beings, Being is prone to be taken for granted,[35] as water is to fish. Nevertheless, “beings give adequate information about Being,” and the Being of beings shows itself as the ground of beings: “For since the beginning of philosophy and with that beginning, the Being of beings has showed itself as the ground. … As the ground, Being brings beings to their actual presencing” (2002, p. 56). Because Being “showed itself as the ground” that brings beings “to their actual presencing,” it is possible to contemplate about Being by philosophizing beings. Metaphysics reflects on what are beyond our immediate senses of the world by “represent[ing] [what is present] in terms of its ground as something grounded” (2002, p. 56). Hence, metaphysics only reasons in representational thought. It “thinks beings as being in the manner of representational thinking which gives reasons” (2002, pp. 55–56). Heidegger gives some examples of such representational thinking about the “ground” in the recent history of metaphysics:

> In accordance with the actual kind of presence, the ground has the character of grounding as the ontic causation of the real, as the transcendental making possible of the objectivity of objects, as the dialectical mediation of the movement of the absolute Spirit, of the historical process of production, as the will to power positing values. (2002, p. 56)

Each of these examples—the instrumentalization of modern sciences, Kantian transcendentalism, Hegel’s dialectic of the Spirit, Nietzsche’s will to power leading to the revaluation of all values—has served as the ground for how beings are revealed. Every epoch can be roughly characterized by such a “ground” or horizon, which determines the presencing of beings and their relations, “the belonging together of beings” (2002, p. 56), including Dasein’s relation to beings.

Heidegger traces the origin of representational thought to the beginning of metaphysics. In “Overcoming Metaphysics” (1973b), he wrote, “[m]etaphysics has distinguished for ages between what beings are and that beings are, or are not” (1973b, p. 2). “What beings are” belongs to the question of essence, and “that beings are”

[35] More accurately, it is the thatness of Being that is taken for granted (Heidegger, 1973a, p. 11).

belongs to the question of existence. Essence "means … that which, for example, the tree as tree, as something growing, living, as treelike, is without any regard to the question whether and that this or that tree 'exists'" (1973b, p. 2). Essence is concerned with "whatness," which "encourages the predominance of looking at what beings are" and is therefore characterized by "the precedence of beings" (1973a, p. 11). Existence is concerned with "thatness," which "establish[es] that beings are" whereby "the essence of Being is assumed as self-evident" (1973a, p. 11). Both "the precedence of beings and the assumed self-evidence of Being, characterize metaphysics" (1973a, p. 11). Being is "the unity of whatness and thatness" (1973a, p. 11).

The connection and distinction between essence and existence can be established historically "with the thinking of Aristotle, who first brought the distinction to a concept" and with "Plato's thinking … that prepared that distinction" (1973b, p. 4). Heidegger argues, this distinction actually marks the beginning of metaphysics: "Being is divided into whatness and thatness. The history of Being as metaphysics begins with this distinction and its preparation" (1973b, p. 2). In other words, Being is not divided into whatness and thatness prior to the beginning of metaphysics. This distinction marks the beginning of metaphysics, and because the determination of this distinction is "an event in the history of Being" (1973b, p. 4) through which "the beginning of metaphysics is revealed" (1973b, p. 2), the distinction is inherent in metaphysics, which "can never of itself come to a knowledge of this distinction" (1973b, p. 3). Throughout the history of western philosophy, the distinction between whatness and thatness, which originates in the beginning of metaphysics, have undergone changing forms: "Throughout the whole history of philosophy, Plato's thinking remains decisive in changing forms. Metaphysics is Platonism" (2002, p. 57). For Heidegger, the changing forms of Plato's thinking witnesses the decline on the unconcealment (*aletheia*) of Being, as Being increasingly withdraws itself into self-concealment (Heidegger, 1973b, pp. 85–86). Over this increasing withdrawal, the unity of whatness and thatness can no longer be recovered, and the extremity of whatness over thatness is manifested in the self-grounding subjectivity of the modern man, to whom "things can only appear as representations (*Vorstellungen*)" (Davis, 2007, p. xxx).

As Bret W. Davis remarks, "[i]t has often been remarked that Heidegger's history of being resembles a kind of inversion of Hegel's history of Spirit. Whereas Hegel sees the movement of history as one of progress toward Spirit's self-realization, Heidegger's history of being would, on this account, tell the story of a decline of the West" (Davis, 2007, p. 157). While the pre-Socratic Greeks "were presumably attuned in wonder amidst the unconcealment (*aletheia*) of being," the abstract and universal concepts in metaphysics, through which beings are represented, gradually curtail the potentiality of beings by making them subservient to the will of representational thought, effecting the withdrawal of Being into self-concealment. For Heidegger, this withdrawal of Being culminates in the epoch of modern technology, "where humans have complete forgotten or been abandoned by being in their frenzy of willful domination of the world" (Davis, 2007, p. 158), where all beings are reduced to "standing-reserves" and exploited as means to technological goals.

Heidegger sees this willful domination of the world through sciences and technology as the final stage of the historical development of metaphysics. He claims that, with the emergence of cybernetics, "philosophy is ending in the present age" (2002, p. 58). "[T]he end" in this context should be understood "a completion" that is "the gathering into the most extreme possibilities" (2002, p. 57). In the most extreme possibilities of whatness over thatness, in an epoch that things can only appear as representations, "philosophy turns into the empirical science of man … the scientific discovery of the individual areas of beings" (2002, p. 57). This "development of philosophy into the independent sciences … is the legitimate completion of philosophy" (2002, p. 58). All these sciences would "soon be determined and guided by the new fundamental science which is called cybernetics" (2002, p. 58). As the scientific attitude takes on this technological character of cybernetics, the "sciences are now taking over as their own task … the ontologies of various regions of being (nature, history, law, art)" (2002, p. 58). Since cybernetics reduces all beings to information, it transduces the management of labour, language, and arts into activities represented in systems and exchanges of information. In this sense, "technology more definitely characterizes and regulates the appearance of the totality of the world and the position of man in it" (2002, p. 58). As a result, "[t]he end of philosophy proves to be the triumph of the manipulable arrangement of a scientific-technological world and of the social order proper to this world" (2002, p. 59). In this final stage of metaphysics, "the operational and model character of representational-calculative thinking becomes dominant" (2002, p. 59).

Ever since Plato, our thoughts have been split into the physical realm and the metaphysical realm. Sciences used to be the branches of knowledge responsible for understanding the physical realm, or the ontical beings. Western philosophy is metaphysics, which attempts to theorize the totality of the world and how Dasein relates to this totality. Metaphysics used to play the role of making sense about the totality of our world. Philosophy over the course of its history has tried to present, in part, "the ontologies of various regions of being," with a coverage that includes nature, history, law, and art. This position of philosophy is now taken up by cybernetics, which is now the new "ground," the technological "clearing," for beings to appear. Within a cybernetic system, all beings can be represented as information. Indeed, if we look at our physical reality today, digitalization has become pervasive in almost every area of living, to the point that digital profiles generated from big data can serve the horizon for human decisions and actions.[36] What Heidegger saw in cybernetics is more than just a powerful methodology for technology innovation, but a "ground" that shapes how people make sense about everything in our world. This view seems remarkably prescient in light of the latest development in technology, from big data and machine learning to the Internet of things.

2.3. The History of Metaphysics as the History of the Will

According to Bret W. Davis, it is possible to see how Heidegger articulates the history of metaphysics as the history of the will by surveying his entire corpus over his lifetime. He traces the path of Heidegger's thoughts on the history of being in *Heidegger and the Will: On the Way to Gelassenheit* (2007). The book points out how Heidegger confronts the problem of the will throughout his works, implicitly in his early writing such as *Being and Time*, and more explicitly in his mature writing since his lectures on Nietzsche's will to power. Davis draws the following conclusion after a comprehensive inquiry of Heidegger's corpus: "Heidegger reads the history of metaphysics as a series of epochs linked together by a narrative of the rise of willful subjectivity, a story that culminates in the technological 'will to will'" (Davis, 2007, p. xxiv). Accordingly, to grasp Heidegger's critique of the essence of modern technology, one needs to first understand his critique of the "technological will to will" and what he means by "will" and "will to will."

In his lectures on Nietzsche's will to power, Heidegger states, "[w]illing itself is mastery over [something], which reaches out beyond itself; will is intrinsically power. And

[36] See Stiegler's concept of tertiary protention (2011, 2016).

power is willing that is constant in itself. Will is power; power is will" (Heidegger, 1991, p. 41). Heidegger rejects the common notion that the will is a faculty of the subject. In the western philosophical tradition, "we tend to understand the will as a 'faculty of the subject,' to be distinguished from 'thinking' or 'feeling'" (2007, p. 5). The will, then, would be on the same level as thinking or feeling. But this is not the will that Heidegger is concerned with. He explains, "[b]y the word 'will' I mean, in fact, not a faculty of the soul, but rather … that wherein the essence of the soul, spirit, reason, love, and life are grounded" (Davis, 2007, p. 6; Heidegger, 2007, p. 78). The will is that wherein being and thinking are grounded. Heidegger directs us to think through "the very ontology which sets up a subject who stands over against a world of objects, to which it then reaches out by means of faculties, powers of representational thought and volitional action, is itself determined by a willful manner of being and thinking" (Davis, 2007, p. 6). His project is "to show that traditional (especially modern) thinking, as representing, is a kind of willing" (Davis, 2007, p. 5) as "[t]o think is to will, and to will is to think" (Heidegger, 1969, p. 59).

According to Davis, the earlier Heidegger sees this will as "the fundamental attunement of the subject who seeks to dominate the world" (Davis, 2007, p. 8), and attunement is the presupposition for our thinking, doing, letting (Davis, 2007, p. 7). A willful fundamental attunement sets up and distorts the horizon for a subject to relate to a world of objects. It "first determines the ontology wherein a subject is open to a world of objects in such a manner that the 'open to' of this relation is distorted (constricted) into the representation of objects present-at-hand" (Davis, 2007, p. 8). The later Heidegger takes this one step further (Davis, 2007, p. 8), claiming that "the being of beings appears as will" to modern metaphysics (1968, pp. 91–92). He contends, "[t]he will in this willing does not mean here a capacity of the human soul . . . ; the word 'willing' here designates the Being of beings as a whole. Every single being and all beings as a whole have their essential powers in and through the will" (Heidegger, 1968, p. 91). As Davis explains, according to Heidegger's "mature being-historical thought[,] … being is revealed-in-(extreme)-concealment as will in the epoch of modernity" (Davis, 2007, p. xxx), and the will is "the name of the being of beings in the epoch of modernity" (2007, p. 8).

To gain a better understanding on how the will, as the being of beings, determines the distortion of the "open-to" relation between a subject and a world of objects, it is necessary to explicate the interconnectedness of representation and will.

Heidegger claims that “thinking, understood in the traditional way, as re-presenting is a kind of willing” (Heidegger, 1969, p. 58). As Davis explains, “Heidegger claims that the Western tradition has in fact reduced knowing to a matter of willing. Knowing has become a matter of subjective representation wherein the world is reduced to the sum total of objects for the representing subject” (2007, p. 174). Instead of letting beings be, that is, letting whatever presencing in the world be themselves, or act and speak for themselves, the western tradition splits beings into the subject-object duality. Knowledge becomes a matter of how objects out there can be represented in the thoughts and languages for the subject. What precisely is representation? “Representation, or as Heidegger sometimes hyphenates the German term, *Vor-stellung*, is a matter of ‘setting before,’ placing an object (*Gegenstand*) in the position of standing over against (*gegen*) the subject, and ultimately at the disposal of his will” (Davis, 2007, p. 174). Representation thus involves the subject reaching out beyond itself, setting the excess before itself, and incorporating the excess back into the domain of the subject. Willing is “being master out beyond oneself” (Heidegger, 1991, p. 63), and “in willing we know ourselves as out beyond ourselves; we sense of mastery over [something], somehow achieved” (Heidegger, 1991, p. 52). Since willing is a matter of the subject “exceeding ourselves only to bring this excess back into the self” (Davis, 2007, p. 9),[37] representation is a matter of placing an object ultimately at the disposal of the subject’s will.

As Davis further explains, representational thought conceals the openness of beings by commanding how beings can appear.

> The subject of representation places himself in the center of beings, or even under them as their ground. Representing is not an open receptiveness to the self-showing of beings, nor is it an ‘engaged letting’ things show themselves from themselves; it is a constitutive knowing (Kant) that commands the very terms in which beings can appear (2007, pp. 174–175).

This is most evident in Kant’s transcendentalism, in which transcendental categories are the necessary conditions for intuiting the objects from our senses. Rather than letting beings be, our representational thoughts, such as Kant’s categories, willfully dictate “the very terms in which beings can appear.” But this will of representation subsists, not only

[37] Davis uses the term “ecstatic-incorporation” to denote this double-sided or “duplicitous” character of will (2007, p. 9).

in Kant, but throughout the history of metaphysics: “The progressive emergence of the will in correlation to the increasing withdrawal of being thus provides a marked aspect of continuity to the history of metaphysics” (Davis, 2007, p. 158). The will to encapsulate the totality of our world into the representational thought of some philosophical system is present throughout the historical progression of western metaphysics, in which representational thought increasingly takes away the openness of being in our world.

The origin of this will can be traced to the Socratic philosophers, as “it is in Aristotle that Heidegger finds the origin of the concept of the will” (Davis, 2007, p. 160).

> Willing is a kind of desiring and striving. The Greeks call it *orexis* . . . But will, as striving, is not blind compulsion. What is desired and striven for is represented as such along with the compulsion . . . What does Aristotle teach concerning the will? The tenth chapter of Book 3 [of *De anima*] deals with *orexis*, desiring. Here Aristotle says (433a 15ff.): “. . . on the basis of [what the desire aims at,] the considering intellect as such determines itself; . . . for what is desired in the desiring moves, and the intellect, representation, moves only because it represents to itself what is desired in the desiring.” . . . (Heidegger, 1991, pp. 54–56)

In Aristotle, the intellect and representational thought moves because of the movement of the will. Heidegger asserts, “Aristotle’s conception of the will becomes definitive for all Western thought” (1991, p. 56).[38] This conception of the will subsists in the Reformation as well as in Descartes, F. W. J. von Schelling, and Hegel. After the Reformation, “*Rectitudo appetitus rationalis*, the correctness of will, the striving for correctness, is the basic form of the will in its willing … The doctrine of justification, and indeed as the question of certainty of salvation, becomes the center of evangelical theology” (Heidegger, 1992b, p. 51). Consequentially, the Reformation brings about a theology centered around “the self-certainty of the *ego cogito*, reducing the religious sphere to the domain of the subject and the will” (Heidegger, 1973c, p. 67). This transformation is completed in Descartes’s *cogito ergo sum*: “Descartes’s metaphysics completes the transformations of truth to certainty, knowing to representation, and the world to an object for the representing subject,” and his metaphysics “marks a decisive point in the

[38] “In the Middle Ages *voluntas* is interpreted as *appetitus intellectuals*, i.e., *orexis dianoētikē*, the desiring which is proper to intellectual representation. For Leibniz *agere*, doing, is *perceptio* and *appetitus* in one; *perceptio* is idea, representation. For Kant the will is that faculty of desire which works according to concepts, which is to say, in such a way that what is willed, as something represented in general, is itself determinative of action” (Heidegger, 1991, p. 56).

rise of the will as ecstatic-incorporation" (Davis, 2007, p. 168).[39] Heidegger then associates the will with the spirit in German Idealism of Schelling and Hegel. In Schelling, the divine will of love is "the force that unites the will of the ground and the will of understanding in their proper ordering" (Davis, 2007, p. 110).[40] In Hegel's *Phenomenology of Spirit*, "absolute subjectivity as absolute self-appearing representation (thinking) is itself the being of beings" (Heidegger, 1993, p. 223). In the absolute subjectivity of the Spirit, "reason is will, here it is reason as representation (idea) that nonetheless decides the beingness of beings" (Heidegger, 1993, p. 223).

The last metaphysician, for Heidegger, is Nietzsche, who "characterizes his philosophy as reversed Platonism" (Heidegger, 2002, p. 57). In Nietzsche, the being of beings is determined by the will to power. "Truths" in western metaphysics are actually values established by the will to power. They are terms that are subjected to the "revaluations of all values." Nietzsche thus put will over reason, body over mind, the sensuous over the supersensuous. There is nothing to "being" in Nietzsche's philosophy "because this concept itself is nothing but an immanently posited value" (Davis, 2007, p. 186). There is only the will to power "for the preservation and enhancement of life" (2007, p. 186). But for Heidegger, "Nietzsche remains within the domain of the terms he inverts" (2007, p. 186). His overturning of metaphysics is still a metaphysics, albeit an inverted one: "The essence of absolute subjectivity first reaches its fulfillment in such inverted empowering of the will. . . . Will to power is therefore both absolute and—because inverted—consummate subjectivity" (Heidegger, 1993, p. 225). Thus rather than liberating from traditional metaphysical oppositions and the will behind them, Nietzsche brings the history of the metaphysics as the progression of the will to its final and ultimate possibility, giving "consummate expression to the descensional progression of the history of metaphysics toward a delimitation of the being of beings as the will to power" (Davis, 2007, p. 186).

---

[39] See footnote 37 on the meaning of "ecstatic-incorporation."

[40] Heidegger elaborates, "[a]s the will of love, spirit is the will to what is set in opposition. This will wills the will of the ground and wills this will of the ground as the counter-will to the will of the understanding. As love, spirit wills the opposing unity of these two wills" (Heidegger, 1985, p. 128).

This metaphysics of the will to power, by bringing the progression of the will to its ultimate possibility, also brings nihilism to its consummated realization. What is nihilism? Davis explains,

> The "essence of nihilism," according to Heidegger, "is the history in which there is nothing to being itself" (Heidegger, 1993, p. 201). Nihilism is precisely the history of an increasing centralization of the subject (and his willful positing of values) who forgets that "beings are thanks to being," that is, thanks to a granting which first opens a clearing for beings to appear. (Davis, 2007, p. 152)

Heidegger points out the irony in Nietzsche's effort to shake free from the western tradition: "It is precisely in the positing of new values from the will to power, by which and through which Nietzsche believes he will overcome nihilism, that nihilism proper first proclaims that there is nothing to being itself, which has now become a value" (Heidegger, 1993, p. 203). As such, Nietzsche's metaphysics "is not an overcoming of nihilism. It is the ultimate entanglement in nihilism" (1993, p. 203). It becomes "the fulfillment of nihilism proper, because it is the metaphysics of the will to power" (1993, p. 204).

2.4. The Technological Will to Will

The metaphysics of will to power, having completed the history of the western metaphysics, "finally reveals itself as the technological will to will" (Davis, 2007, p. 146). Willing is constantly associated with an insatiability to be more: "Every willing is a willing to be more. Power itself only is inasmuch as, and as long as, it remains a willing to be more power" (Heidegger, 1991, p. 60). The will is "insatiable, ever expanding" (Davis, 2007, p. 11). The "will to will" refers to a will that constantly expands its control as it reaches out to the world in order to reduce this world to the domain of the will. It is the extreme form of the will to power, and what it strives after is not something that it lacks and desires, but simply an expansion of the domain in which the will commands.

> The will is not a desiring, and not a mere striving after something, but rather, willing is in itself a commanding … Commanding, which is to be sharply distinguished from the mere ordering about of others, is self-conquest and is more difficult than obeying. … What the will wills it does not merely strive after as something it does not yet have. What the will wills it has already. For the will wills its will. Its will is what it has willed. The will wills itself. (Heidegger, 1977b, p. 77).

In short, “will to power is will to will, which is to say, willing is a self-willing” (Heidegger, 1991, p. 37). Heidegger further contends, this will to will is the essence of technology. Initially, “in the rise of the modern metaphysics of will things are progressively reduced to objects of human will and representation” (Davis, 2007, p. 178). But eventually, “in the extreme epoch of technology even this egocentric dualism threatens to give way to a uniform ordering about of both non-human and human beings as standing-reserve” (2007, p. 178). The technological will to will expands its domain of commanding over non-human and human beings alike. It “ultimately threatens to strip humans of their freedom, reducing them to another cog in the wheel of machination” (Davis, 2007, p. 121).

Commanding over the human and non-human world becomes possible with cybernetics, a fundamental science that shifts the paradigm of all physical and human sciences to that of information. As Heidegger recognizes, “[i]n the cybernetically represented world, the difference between automatic machines and living things disappears. It becomes neutralized by the undifferentiated process of information” (Heidegger, 1983, p. 142; Davis, 2007, p. 178). Every aspect of a human or non-human being can be represented as information, which is subjected to calculability and control. The cybernetics movement is a method that “makes possible a completely uniform and in this sense universal calculability, in other words the controllability of the lifeless and the living world. In this uniformity of the cybernetic world, man too gets installed [*eingewiesen*]” (Heidegger, 1983, p. 142; Davis, 2007, p. 178). Hence, cybernetics can be regarded as the “extreme form of modern technology” and the ultimate expression for the technological will to will (Davis, 2007, p. 178):

For Michael E. Zimmerman (2016, p. 106), Heidegger’s late reflections on cybernetics were prescient as the technological will to will anticipates the techno-posthumanist construct of the super AI: “In coming decades, so Heidegger surmised, the Will to Power will allow and even demand that humans generate what today is depicted as autonomous, super AI” (2016, p. 106). Zimmerman adopts the terms transhumanism and techno-posthumanism from Ray Kurzweil (2005).[41] The central goal of transhumanism or transitional humanism is enhancing the human (Zimmerman, 2016, p.

[41] Note that techno-posthumanism is a form of posthumanism that points to something very different from the philosophical-literary posthumanism developed by thinkers such as Jacques Derrida and Michel Foucault (Zimmerman, p. 98-99).

98). Kurzweil predicts that transhumanism will progress toward techno-posthumanism in the not-too-distant future, when “transhumans will merge with super AI, which may evoke from ordinary humans the awe formerly associated with encountering the gods” (Zimmerman, 2016, p. 98). In Kurzweil’s futuristic vision, the human species will transition into an evolving humankind equipped with possibilities and capabilities that were traditionally accorded only to a deity (Zimmerman, 2016, p. 98). Thus virtual immortality, omniscience, or mastery over nature are all recurrent themes in the discourse on techno-posthumanism (2016, p. 98). In fact, a “central motivation for transhumanism and techno-posthumanism is to avoid death” (2016, p. 98). According to the predictions by Kurzweil and other techno-posthumanists, within relatively short time, AI will be able to redesign itself such that “AI will be billions of times more intelligent than human beings” (Zimmerman, 2016, p. 98). What Kurzweil calls the Singularity is the moment when AI surpasses human intelligence.

Zimmerman develops his critique of transhumanism and techno-posthumanism by drawing from Heidegger’s writing on the essence of modern technology and the technological will to will. He tries to project Heidegger’s perspective on techno-posthumanism as if he were still alive today:

> Heidegger would say that techno-posthumanism is the latest and perhaps most dangerous phase in the era of techno-industrial nihilism. In this era, all beings—including human beings—reveal themselves primarily as raw material for the purpose of enhancing power for its own sake, not for the sake of some identifiable human end. (2016, p. 101)

Techno-posthumanism could be the most dangerous phase of modern technology, whose essence is the enframing of all beings under the technological will to will. This phase began with cybernetics, which “seems to be late-Heidegger’s operational term for ‘enframing’ (*Gestell*)” (2016, p. 102). Cybernetics is destined to bring about the materialization of the essence of technology, the technological will to will, which ultimately finds its nihilistic materialization in the super AI: “super AI would continually enhance itself, becoming ever-more powerful, perhaps as an end in itself. Super AI would be, in effect, the ultimate ontical embodiment of what Heidegger—drawing on Nietzsche—calls the Will to Will” (Zimmerman, 2016, p. 101). It follows that “[t]he metaphysics of the Will to Will is discernible in Kurzweil’s prediction that an artificial intelligent computer will redesign itself, so that it can become far more intelligent than *all human beings collectively*” (2016, p. 104).

As Davis comments, “technology no longer centers on the striving power of the self-interested subject (Davis, 2007, p. 179). The essence of modern technology, for Heidegger, is revealed in the insatiable expansion of the technological system to command over both the lifeless and the living world. In the age of cybernetics, this essence mobilizes nearly all contemporary technical activities. As we will see in later chapters, this totalizing critique on the universal calculability of cybernetics is very influential on the thoughts of Stiegler and Yuk Hui. The abstract critique of the essence of technology, of the technological will to will, also contradicts the Marxist critique that centers on the striving of power by the capitalist class. We will come back to this contradiction in Chapter 8 and deliberate whether Marcuse’s technological rationality, on which Feenberg’s technical politics is based, is a sufficient Marxist response to the age of cybernetics.

2.5. Heidegger’s Eschatology

One common criticism that Heidegger faces over his critique of modern technology concerns its implied notion that humanity is helpless against the domination of technological system, to the point that “only a god can save us.” But according to Davis, this criticism is not entirely accurate, as Heidegger has in fact laid a path forward for others to follow in their intellectual pursuits.[42] Drawing analogy from Christian eschatology, Heidegger makes conjectures about a new other beginning after which the world and its technology will develop after breaking free from the domain of the will. Paradoxically, this new other beginning is only possible after humanity has reached the extreme epoch of the technological will to will:

> [It] is precisely in this extreme danger that the innermost indestructible belongingness of man within granting may come to light, provided that we, for our part, begin to pay heed to the essence of technology.
>
> Thus the essential unfolding of technology harbors in itself what we least suspect, the possible rise of the saving power. (Heidegger, 1953, p. 337)

This move follows the Christian narrative about the Second Coming of Jesus Christ: The coming to an end of our world brings forth the new Heavenly Kingdom. It is also

[42] Among the notable ones are Herbert Dreyfus (1965, 1972, 1992), Albert Borgmann (1988), and Bernard Stiegler.

reminiscent of the Marxist theory about the inevitable collapse of capitalism when the proletariat becomes conscious of the alienated condition of their class, prompting resistance and rebellion. Heidegger's eschatology of salvation from technological domination results neither from divine intervention nor from class consciousness. Rather, it involves an epochal change of collective awareness about the true condition of humanity, that is, the openness of its lifeworld has been taken away by its increasing dependence on technology. Thus "the danger is, when seen as the danger, at once the growth of that-which-saves" (Davis, 2007, p. 182). This epochal change can be partially attributed to human actions or political projects, but is in general unpredictable as to how and when it may come about.

To see how this epochal change of a new other beginning is possible for Heidegger, we must recognize his differentiation between technology and the essence of technology. Heidegger is well aware that technology and our lifeworld are inseparable, that technology as human-fabricated nature makes up the very texture of our lifeworld and has always been defining for what it means to be human. Thus he wrote,

> It would be foolish to attack technology blindly. It would be shortsighted to want to condemn it as the work of the devil. We depend on technical devices; they even challenge us to ever greater advances. But suddenly and unaware we find ourselves so firmly shackled to these devices that we fall into bondage to them. (Heidegger, 1969, pp. 53–54)

What needs to be addressed is not technology per se, but the relation of bondage that creeps into our dealing with technology. This is what he means in the following passage from "The Question Concerning Technology" (1953): "What is dangerous is not technology. Technology is not demonic. . . . The essence of technology, as a destining of revealing, is the danger" (Heidegger, 1953, p. 333). Therefore, "Heidegger's critique is not aimed at technological devices themselves, but at the way of revealing/concealing which they embody" (Davis, 2007, p. 183). Technological devices embody the way all beings are revealed as standing-reserves for technology's ever expanding will of mastery, a will that is not tangible but nonetheless manifests itself through the beings in the modern epoch. Therefore, the devices in themselves are not necessarily "demonic" and may possibly be salvaged from the technological will to will.

To prevent falling into bondage to technology, Heidegger proposes a proper use of technical devices such that we can keep ourselves free from them, a particular technological mentality of letting beings be. This way, technology would not exhaust natural resources as standing-reserves for the ever-expanding technological will to will, nor would it affect our most inner core as human beings:

> We can use technical devices, and yet with proper use also keep ourselves free of them, so that we may let go of them at any time . . . let them alone as something which does not affect our most inner and proper [essence]. . . . We let technical devices enter our daily life, and at the same time leave them outside. . . . I would call this comportment toward technology, which expresses "yes" at the same time as "no," by an old word: releasement toward things [*die Gelassenheit zu den Dingen*]. (Heidegger, 1969, p. 54, 1992a, pp. 22–23)

This comportment toward technology, *die Gelassenheit zu den Dingen*, "is not only a releasement from willful technological manipulation; it is also a releasement into a more attentive engagement in letting things be" (Davis, 2007, p. 184). For Heidegger, the epochal change of salvation from technological domination involves a collective turn in humanity toward *Gelassenheit*. This collective turn would subvert the perpetual concealment of being over the history of the will, leading to an epoch of openness of letting beings be.

In Heidegger's theory, even though this collective turn in humanity is not realizable purely through humans' actions, humans can nonetheless play a role in bringing upon a new beginning of the post-metaphysical West. The role they play is to reveal the concealed essence of modern technology (an endeavour that Heidegger is leading the charge) and subsequently to adopt a attunement of non-willing [*Nicht-Wollen*] toward being. The non-willing comportments of "releasement towards things [*Gelassenheit zu den Dinged*] and openness to the mystery … do not simply befall us accidentally [*Sie sind nichts Zu-fälliges*]," but rather "through "persistent, courageous thinking [*einem unablässigen herzhaften Denken*]" (Heidegger, 1969, p. 56, 1992a, p. 25). *Gelassenheit* is a concept adopted from the medieval Christian mystic and German theologian Meister Eckhart. Roughly speaking, Heidegger adopts the Christian tradition of deferring the self-will to the Divine Will. But deferred willing still belongs to the domain of willing. Moreover, deferred-willing can easily be turned into a kind of covert-willing in which spokespersons for the divinities would exhibit their will to power. Thus the late-Heidegger is concerned with weaning from any trace of willing in order to subvert the will to power and the will to will as the Being of beings in the modern epoch. He contemplates on a non-willing [*Nicht-Wollen*] comportment toward all beings. Heidegger defines non-willing as follows:

> *Nicht-Wollen* means, accordingly: [1] willing to renounce willing [*willentlich dem Wollen absagen*]. And the term means, further: [2] what remains strictly outside any kind of will [*was schlechthin ausserhalb jeder Art von Willen bleibt*]. (Heidegger, 1969, p. 59, 1992a, p. 30)

To move toward a comportment of non-willing, humanity first needs to assertively renounce willing in all their activities, but this assertiveness is also a kind of willing. So we need to take this one step further by entering into a non-willing state in which we simply live, think, and act in a mode of openness, without willing and without reacting against willful domination. This second step seems paradoxical and is most tricky, and Davis suggests that it is ultimately only "by way of a leap (*Sprung*) that one could move from the domain of the will into the region of non-willing, even if a painstaking 'twisting free' is needed to prepare for this leap" (2007, p. 188). He further elaborates, "[n]on willing, in the most radical sense of other than willing, could only be reached by way of undergoing an arduous twisting through a paradoxical willing non-willing. *Nicht-Wollen* is for us today, if not always, inherently ambiguous. It is both an ascetic weaning from, and a way of being other than, willing" (2007, p. 203).

2.6. Conclusion

In this chapter, we began with Heidegger's remarks in *Der Spiegel* magazine that cybernetics would take the place of philosophy as the horizon of effecting changes in the western world, and that only a god can save humanity from the total domination of technology. With the help of the commentaries by Davis and Zimmerman, we proceeded to explain Heidegger's remarks by surveying his corpus over his lifetime. We discussed how the problem of the will is central to Heidegger's critique of the essence of modern technology. This essence is the *Ge-stell*, typically translated as "enframing", as *stell* in German means frame or position. In Davis' exegesis, the "'positing' (*Setzen*) or 'positioning' (*Stellen*) character of the will, which represents (*vor-stellt*) its objects as means to its own securing and enhancing of power, is at the heart of what Heidegger problematizes as the *Ge-stell* of technology" (Davis, 2007, p. 151). Within this technological *Ge-stell*, everything, both the living and the non-living, is posited, positioned, and represented in accordance to an abstract, invisible will in the insatiable expansion of technology. Cybernetics is the ultimate ontical embodiment of this essence. This fundamental science would supersede the traditional role played by philosophy in its provision of a ground, a horizon, for determining how beings are revealed. Dasein, being thrown into a world of beings, can only make sense of this world and the beings-in-relations by associating them with meanings based on some representational system. Philosophy and religion used to be man's attempts to come up with representational systems to grant meanings and sanity to people's lives. These attempts in the West has yielded a history of metaphysics that progressively conceals beings, which become increasingly subservient to will to power and ultimately to the technological will to will. As Zimmerman points out, the futuristic imagination of super AI and Singularity, proposed by techno-posthumanists such as Kurzweil, could become the most dangerous phase in the epoch of technological will to will.

This ontological critique of modern technology seems to suggest that humanity is helpless within technological *Gestell*. Any attempt to shake free from the technological will to will would paradoxically come back as the new dominating will, as in the case of Nietzsche's will to power. Heidegger seems to confirm this helpless sentiment with his comment that "only a god can save us." But as Davis shows us in his explication of *Gelassenheit*, Heidegger did point to a path forward by articulating a fundamental attunement of "releasement towards things" [*Gelassenheit zu den Dinged*]. So while humanity cannot, by its own will, bring upon a new beginning of non-willing, this new epoch actually becomes possible when humanity enters into the abyss of technological nihilism. Finding themselves in this abyss would bring about an awakening to people on how they are becoming standing-reserves for the sake of technology. The emanation of such awakening signals a new epoch of non-willing. While we can only wait for this awakening and let this happen historically, there is hope that the awakening may be happening already, presumably in Heidegger's own philosophy but also in other non-positivistic critiques since Heidegger. Thus Heidegger's statement that "only a god can save us" can be interpreted as a hopeful rather than a hopeless vision of our future world.

Yet, critics of Heidegger find his spiritual approach overly passive and abstract.[43] Can we formulate a critique of the current state of technology, in particular big data and machine learning, that suggests a more down-to-earth approach in steering the direction of research and development for these technologies? This entails the possibility of a "new science" that Marcuse proposes and Feenberg further develops in his technical politics. There may be an unresolvable tension between the assertiveness of technical politics and the passiveness of Heidegger's non-willing *Gelassenheit*, but they do share the common goal of letting things be. In their critiques of technology, both Marcuse and Feenberg refer to Gilbert Simondon, a philosopher who refutes a dystopian critique of technology and one who deliberates on how humans can engage in a positive relationship with technology.[44] Simondon's philosophy is also of particular relevance to our discussion because it was developed as a response to the cybernetics movement. This dissertation deliberates on the significance of cybernetics to Simondon's philosophy and the implication of his philosophy on artificial intelligence and machine learning from Chapter 6 to Chapter 9. But before engaging in this philosophical inquiry, it would be helpful to first look into how cybernetics is related to artificial intelligence (Chapter 3) and to machine learning (Chapter 4), as well as to examine the potential affordance and limitation of machine learning and deep learning (Chapter 5).

[43] Feenberg talks about such critics, including Marcuse, in *Transforming Technology* (2002, p. 7) and in *The Ruthless Critique of Everything Existing* (2023, p. 28). Feenberg himself is sceptical of Heidegger's approach as it provides no pathway to an emancipating technical politics. For more, see Feenberg's *Heidegger and Marcuse* (2005, pp. 21–45), *Technosystem* (2017a, pp. 140–148).

[44] One key question for forging such a positive relationship is where the potentiality for changes may originate from. Heidegger attempts to overcome western metaphysics in his critique of the will. But this is not to say that he rejects any notion of reality beyond the empirical. He acknowledges both an essence of technology and an essence of man, which is "the innermost indestructible belongingness of man within granting" (Heidegger, 1953, p. 337). This essence of man is not reducible to an alienated will, and "[i]t is precisely this intimation of an 'innermost,' 'indestructible' and 'free essence' of man, which exceeds the modern historical delimitation of his essence as will, that would enable a human participation in the transition to an other beginning of non-willing" (Davis, 2007, p. 213). Such writing on "essence" reflects a non-empirical notion of potentiality. In this sense, Heidegger's essence is comparable to Simondon's theory of pre-individuality. As I will explain in Chapter 7, Simondon formulates his theory of pre-individual reality by identifying potentiality in the obscure zones of reality that is intrinsically beyond the deterministic predictions of empirical sciences.

Chapter 3.

The Genealogy of AI: Questioning the Universality of Cybernetics

Me what do people mean by the universality of cybernetics?

ChatGPT: The universality of cybernetics refers to the idea that the principles and concepts of cybernetics can be applied to a wide range of systems, across different disciplines and domains. Cybernetics is concerned with the study of control and communication in systems, whether they are mechanical, biological, social, or cognitive. This broad focus allows cybernetics to be applicable to a diverse range of systems, from individual organisms to complex organizations.

The universality of cybernetics is often attributed to its interdisciplinary nature. Cybernetics draws on concepts and methods from a variety of fields, including engineering, mathematics, biology, psychology, and sociology. This cross-disciplinary approach allows cybernetics to provide a framework for understanding and designing systems that involve feedback, control, and communication.The "will to power" is a central concept in the philosophy of Friedrich Nietzsche, who describes it as the basic drive behind all human behavior, thought, and action. It refers to the instinctual urge to increase one's own power and control over one's environment, both internally through self-mastery, and externally through domination over others. The "will to power" is not just a desire for power in a conventional sense, but a deep-seated striving for excellence and self-overcoming. It is a fundamental aspect of the human condition and, according to Nietzsche, an essential part of life and growth.

3.1. Overview

In the previous chapter, I present Heidegger's critique of cybernetics as the completion of metaphysics and culmination of the technological will to will. He sees cybernetics as a fundamental science that grounds all other sciences, with all beings reduced to systems and exchanges of information, subject to the control of human wills or some abstract technological will. This critique on the universality of cybernetics has indeed been taken up by others (see Bowker, 2008; Dupuy, 2000; Hayles, 1999; Pias, 2003b). Characterizing cybernetics as universal in its control of all beings makes it an easy target for social criticism, but is this characterization true to the historical development of cybernetics and the implication it has left us with? Much has been written about the universality of the cybernetics movement (e.g. Bowker, 2008; Hayles, 1999; Hui, 2019), but social and philosophical critiques have a tendency of taking a bird's-eye view without getting into the nitty-gritty of the actual technical discussions. To get a true sense about what went on in the movement, it would be most direct to examine the actual conversations that took place between the scientists who participated in the Macy Cybernetics Conferences.

The Macy Cybernetics Conferences, held from 1946 to 1953, was the the marquee event for the cybernetics movement. A few key themes in these cybernetics meetings were intriguing to the leading experts across different disciplines. These themes are (1) negative feedbacks that makes possible controls in a dynamic environment, (2) the possibility to represent all living and non-living beings as information, and (3) the blurring of boundaries between the human and the non-human, between the living and non-living. Participants included top neurologists, psychiatrists, biologists, mathematicians, physicists as well as early pioneers of computing such as John von Neumann and Claude Shannon. Nevertheless, if we examine the actual conversations at the conferences, the exchanges among the participants were filled with strife and disagreement due to the inherent contradictions between long-held assumptions within each field of knowledge. Many quit in frustration. These anecdotal accounts show that scientific research in reality cannot be reduced to some clean and simple theory of universality, thus problematizing the social and philosophical critique of cybernetics as a universal discipline of all disciplines.[45]

As Jean-Pierre Dupuy remarked, far from being a stunning success, "cybernetics ended in failure. It was a historical failure, one that was all the more bitter as its advertised ambitions were enormous" (2000, p. 15).[46] In retrospect, it is difficult to gauge

---

[45] In this problematization of the critique of universality in cybernetics, I am attempting to engage the critical literature on cybernetics (from Chapter 2) with the perspective of technological research, resulting in a dialectical analysis that is both critical and immanent. This dialectical approach will resurface later when I discuss Simondon's philosophy (Chapter 6 to Chapter 8) and the social implication of machine learning (Chapter 9).

[46] Dupuy explains, "[t]hose who dedicate themselves to this purpose today may find it useful to immerse themselves once again in these pioneering debates. If any further reason is needed to

the success of the Macy Conferences because they are simply meetup events that gathered together researchers who had been working on independent research projects across different knowledge domains. Nevertheless, the interdisciplinary effort did facilitate exchanges between different fields. The overlapping of multi-disciplinary knowledge could serve as the impetus for new research ideas. As I will show, the birth of AI can be attributed to the clash of ideas between the mainstream research of computer as calculating machines and the cybernetic idea of blurring the boundary between man and machines.

In this chapter, I begin by digging into the published transactions of the Macy Cybernetics Conferences and other historical documents. The disjointed conversations in these transactions cast doubt on the critical notion that universality characterizes cybernetics. These conference transactions, which give a glimpse into the cybernetics movement, suggests irresolvable tensions and contradictions between respective disciplinary assumptions. Yet, as I will argue, it is precisely the clashes of seemingly unrelated knowledge domains that brought forth innovations in various knowledge fields. Thus, after questioning the universality of cybernetics, I will trace the historical origin of the computer from the universal Turing machine (UTM) to the Analytic Machine designed for mass calculation by Charles Babbage in the 19th century. I then elaborate on how ideas in cybernetics have been influential to the historical appropriation of the calculating computing machine into modern-day computing with many other purposes. Of particular relevance to this dissertation is the historical emergence of AI, which can be seen as a crossbreeding between the origination of computers as calculating machines with the cybernetic theme of boundary-crossing between humans and machines. I will describe the early history of AI to see how the field branched out of research in cybernetics, and explain the adoption of both ideas in cybernetics and the design principles of computer sciences in AI research.

---

convince them of this, it would be the following, which is only apparently paradoxical: cybernetics ended in failure. It was a historical failure, one that was all the more bitter as its advertised ambitions were enormous; a conceptual failure, all the less comprehensible in view of the fact that it had marshaled very great intellectual advantages on its side; and, finally, if we consider all that is owed to it and that has gone unacknowledged, it was perhaps an unjust failure as well" (2000, p. 15).

3.2. Questioning the Universality of Cybernetics

Cybernetician Gregory Bateson once claim that “cybernetics is the biggest bite out of the fruit of the Tree of Knowledge that mankind has taken in the last 2000 years” (1987, p. 481). But this view on the vast implication of cybernetics on the development of science and technology is far from unanimous. Warren McCulloch, who served as the chair for most of the conferences, is more reserved in his judgement of cybernetics. He seems to echo Bateson’s “Tree of Knowledge” appraisal, noting that cybernetics “has been a challenge to logic and to mathematics, an inspiration to neurophysiology and to the theory of automata, including artificial intelligence, and bionics or robotology” (1974/2004, p. 360), but he also surmised that “[t]o the social sciences it is still mere suspiration” (1974/2004, p. 360). In fact, being a challenge and an inspiration to many fields seems to be a far cry from the “biggest bite of the fruit of the Tree of Knowledge” over the last two thousand years. Is cybernetics truly a fundamental science that grounds all sciences, or merely an inspiration to the development of many fields? As I will contend, the narrative about the universality of cybernetics is a problematic discourse from a technical perspective. Thus attributing the totality of a technocratic worldview to cybernetics is a philosophical or social proposition that understates the actual conflicts between disciplinary knowledge and assumptions are effaced. Nonetheless, key ideas from the cybernetics movement have turned out to be influential to the continual development in computing research.

There are many who ascribe the character of universality to the cybernetics movement. The Macy conferences “began chiefly because Norbert Wiener and his friends in mathematics, communication engineering, and physiology, had shown the applicability of the notions of inverse feedback to all problems of regulation, homeostasis, and goal directed activity from steam engines to human societies” (McCulloch, 1953/2003b, p. 719). Geoffrey C. Bowker identified an ideal of universality behind this interdisciplinary mandate of cybernetics. The “cybernetic thesis that purpose could be formed in machines by feedback” unifies physical, biological, chemical, and social sciences, and this ideal of universality devours all disciplines in sciences and humanities into its own abstract form, reducing the vast and complex material reality into a controllable abstraction (Bowker, 2008, p. 77). Much has been written on this universality of the cybernetics thesis (Bowker, 2008; Dupuy, 2000; Hayles, 1999; Pias, 2003b).[47] In Claude Pias' review on cybernetics, this widespread applicability led to the goal of formulating a general theory based on "the principles of the current computer generation, the latest developments of neurophysiology, and finally a vague 'humanistic' combination of psychiatry, anthropology, and sociology" (Pias, 2003b, p. 11). The American research conducted in the early 1940s—Pitts and McCulloch's artificial neuron, Shannon's information theory, and the behavioral theory of feedback by Wiener, Bigelow, and Rosenblueth—were combined at the Macy conferences into a single, "universal theory of digital machines, a stochastic theory of the symbolic, and a non-deterministic yet teleological theory of feedback … that could then claim validity for living organisms as well as machines, for economic theory" (Pias, 2003b, p. 15). This way of framing cybernetics turns it into an agent for technocratic totality, which fits nicely into a dystopian discourse about technology, from Heidegger's enframing and Herbert Marcuse's technological rationality to Ray Kurzweil's technological singularity. When cybernetics becomes synonymous with a science-centric worldview that advocates for the use of scientific methodology in studying humans and societies, scholars such as Bernard Dionysius Geoghegan (2023) can attribute scientific administration and technocratic reforms to cybernetics. But by taking a closer look at the Macy conferences, this framing of cybernetics as an ideology of universality would appear contradictory to the interdisciplinary conflicts in cybernetics, which become apparent when we examine the field from a technical perspective.

It is widely recognized that the two most influential leaders of the cybernetics movement were Norbert Wiener and Warren McCulloch. Wiener was the mercurial, charismatic genius. He was well known for his bizarre practice of jumping into a discussion with an ingenious comment while he had been asleep and snoring throughout (Kubie, 1951/2003, p. 416; McCulloch, 1974/2004, p. 355). After the publication of his best seller *Cybernetics* in 1948, he only sparingly attended the Macy conferences and gradually faded out from the interdisciplinary activities (McCulloch, 1974/2004, pp. 358–359). McCulloch was the conference chair who hosted all the meetings and steadfastly smoothed out disputes among representatives from different backgrounds. There were mainly three groups of people: one group consisted of the physicists and mathematicians; the second group was made up of anthropologists and psychiatrists;

[47] For Dupuy, it is the mechanization of the mind; for Hayles, it is posthuman—the reconceptualization of what the human is.

the third group were "simple people who had a lot of loose intuition and no discipline to what they were doing" (Brand, 1976/2004, p. 303). Chairing meetings with such diverse backgrounds was no easy feat. A number of years later, McCulloch gave a vivid description on the intensity in those meetings: "The first five meetings were intolerable. Some participants left in tears, never to return. We tried some sessions with and some without recording, but nothing was printable. The smoke, the noise, the smell of battle are not printable" (McCulloch, 1974/2004, p. 356).

Participants from various fields had difficulties relating with each other because each field upheld a unique set of implicit assumptions. The discrepancies in these assumptions would lead to clashes in arguments. For instance, Bigelow expressed that "[t]here has been quite a tendency in these meetings to assume that the mathematical methods of the physical sciences are necessarily those appropriate for most of the other sciences and other fields, but there are cases where this may not be true" (Kubie, 1952/2003, p. 591). McCulloch indicated that "this group has been guilty of a certain irreverence with respect to the subconscious or the unconscious" (McCulloch, 1953/2003a, p. 687). In one of the rare instances where members from different backgrounds managed to engage in a fruitful exchange, the medical director of the Josiah Macy Jr. Foundation Frank Fremont-Smith remarked: "This is the thing I have been waiting for since I started this conference group: that we who think in mathematical, physical, and engineering terms would come to grips in a genuine way with the people who think and talk in symbolic unconscious terms. There is a tendency to intolerance on both sides which should be avoided" (Kubie, 1950/2003, p. 318).

The discussions in the conferences were loosely tied together by some common themes or keywords. The published transactions are interspersed with terms like "machinery," "automata," "human machine," "code," "symbols," or "abstractions." Nonetheless, members of each field appropriated the meanings of these keywords to fit their own implicit assumptions. For instance, many variations of feedbacks were discussed throughout the conferences. In some instances, the conceptualizations could be quite remote from the type of feedbacks as exemplified by Wiener's guided missiles. Heinz von Forester discussed the presence of "feedback between writing and speaking" (Shannon, 1950/2003, p. 262). Jerome B. Wiesner identified feedback in how people react to humor: "One does not laugh hard where there is not the possibility of feedback. If you are listening to the radio by yourself or reading a book, you will chuckle, whereas the same stimulus, in a group, may evoke enormous laughter" (Bateson, 1952/2003, p. 557). It almost seems as if participants felt pressured to adopt some notion of feedback into their works in order to fit in the general discourse of the conferences.

Throughout the conferences, different terms always seemed open for interpretations and might cause confusion or miscommunication. Participants often felt confused if "analogue" meant "analogy" or if it meant the opposite of "digital" in a given context. There were heated discussions on vastly different classes of symbols that share common attributes: "symbols" in linguistics, "symbols" of the unconscious in psychoanalytic, and "symbols" in the abstraction of computing machines. Some of the most intense dialogues involved how to define "information" and "communication." The probabilistic and statistical definitions in Shannon's information theory were foreign to psychiatrists and biologists. In one of his presentations, Herbert G. Birch gave his version of "communication" from biology, placing much emphasis on anticipation and expectation: "[T]rue communication, in my sense at any rate – not necessarily in the sense of the communications engineer or in the sense of the physicist, but in the sense of a student of animal behavior – would be represented by this kind of level of interdependent communication that has direction, that involves the process of anticipation, and that involves the process of conventionalization of sound" (Birch, 1951/2003, p. 464). Birch illustrated such communications between organisms by describing a scallop's reaction when a starfish being placed near it. Due to some chemical stimuli coming from the starfish, the scallop would immediately flee. In such examples, "the organism that is now sending has an expectation. It anticipates a behavior upon the part of the other organism" (1951/2003, p. 464).[48] In general, it was nearly impossible for conference participants to standardize the definitions of cybernetic terms.

Coming into consensus on how to define a single term posted enough of a challenge, let alone a unified, general theory across all knowledge disciplines. In fact, they were so far away from a unified theory that even the umbrella term, "cybernetics," was filled with ambivalent meanings. As one of the conference participants blurted out: "Cybernetics is a term that means all things to some men and nothing to many" (Grey-

[48] As I will explain later on, this biological view on information has much affinity with Simondon's theory of information.

Walter, 1953/2003, p. 689). Despite the difficulties to keep everyone in sync, conference participants nevertheless found their involvement very fruitful. One of them, Margaret Mead, felt that "it was the most interesting conference" she had ever attended because "nobody knew how to manage this [*sic*] things yet" (Brand, 1976/2004, p. 303). Without coming into agreement on basic assumptions, ideas from other fields could still be inspirational and adaptable to one's own research. As Heinz von Foerster reflected on his experience in the Macy conferences,

> the thing that is shared is not simply a belief that the different disciplines ought to understand each other better, nor a body of shared material to which different methods of analysis are brought together, nor a single problem towards the solution of which the members are bending their differentiated and united efforts, but rather … an experiment with a set of conceptual models which seem to be useful right across the board and which themselves provide a medium of communication also – when shared. (Pias, 2003b, pp. 14–15)

Cybernetics was, from its very beginnings, "less a disciplinary science than a general methodology of action" (Pias, 2003b, p. 23).

It should be evident by now that there is no concise definition of cybernetics, and the interdisciplinary effort and communication seem neither perfect nor impactful. Is cybernetics then simply a buzzword for capturing all the transformational development in science and technology since the middle of the 20th century? Did cybernetics leave meagre influences on other knowledge disciplines, or was its influence of any significance? As I will try to show next, the significance of cybernetics comes in the new direction it provides for scientific research, and particularly research in AI and computing. Prior to cybernetics, technics is primarily concerned with the provision of tools or instruments, whose goals are to assist humans in their physical, mental, or social activities. In cybernetics, machines are given the same ontological status as living and organic beings, and technical research under its influence has been steered toward the design of complex systems and experimentation on relating such systems to human users. This paradigm shift is evident in the early development of computers, in which the original dominant image of a calculating machine gave way to the image of an intelligent machine.

3.3. The Origin of Computers as Calculating Machine[49]

Looking up the etymology of "computer," this word was originally reserved for people who helped the government undertake the preparation of mathematical table in early 1800s (Bowden, 1953, p. 247).[50] A hundred and fifty years later, what we call computers today were more often referred to as "calculating machines," "computing machines," "digital computers," or "automatic computers" in order to distinguish them from "human computers." In *Faster than Thoughts* (1953), B. V. Bowden often qualified computers with "humans" or "digital" to distinguish between the two. For instance, he would write, "[a] human computer working at his desk needs a calculating machine, reference books of tables, pen and paper with which to record the intermediate results in his calculations, and instructions as to how to proceed" (1953, p. 26). Here, "a human computer" is the person whereas the "calculating machine" may refer to a computer or a calculator. When he neglected to add the qualifier, readers today may find it confusing. That is the case when he wrote "hiring a fast computer" (1953, p. 232). He meant hiring a person, but readers today would first think of a machine before re-interpreting "computer" as a person in order to make sense of the word "hiring."

In this context, the original conception of a universal computer was designed for automating the process of mass calculation. The first universal computer was conceptualized by Charles Babbage. He proposed a mechanical general-purpose computer that "would be able to perform any calculation whatsoever" (Bowden, 1953, p. 10). But he never managed to construct this machine. It was not until around a hundred years later when Alan Turing among others were able to design and build an actual general-purpose computer. The early calculating machines were huge, complicated machines, as was the case with the E.N.I.A.C. built at the Moore School of Electrical Engineering in the University of Pennsylvania in 1946. Bowden provides a snapshot of what computing machines signified to the general public in the 1950s: "During the last year or two most people must have heard of the remarkable devices often called

[49] The purpose of this section is to highlight the original conception of a computer as a calculating machine. Due to the universality of the Turing machine, the conception of a computer evolves throughout its history. For more, see historical accounts on the evolving hardware in its early days (Aspray, 1990; Mindell, 2002) on the transition into an information machine (Campbell-Kelly & Aspray, 2004), and on the transition into a personal device under the influence of cybernetics (Bardini, 2000; e.g., Bowker, 2008; Turner, 2008). See also further discussion in Section 3.4.

[50] For more, see *When Computers Were Humans* (2005) by David Alan Grier.

'Electronic Brains'" (1953, p. vii), and these high-speed electronic brains "were all designed originally to solve scientific problems; they will do as much arithmetic in a week as most men can do in a lifetime" (1953, p. 246). Apparently in the 1950s, the public perceptions of a calculating machine was an electronic brain that replaces human brain for performing arithmetic and scientific computations in high speed (hence the book's title *Faster than Thoughts*).

In Bowden's view, the promise of computing machines came from their potential to improve efficiency in numerous social activities. Computing machines were seen as the harbingers of "a second Industrial Revolution" (1953, p. x), as superior upgrades over the automata in factory assembly line. Instead of replacing manual chores, computing machines were on the verge of supplanting works that used to be performed by the human mind. In addition to playing an important role in scientific research (1953, p. 130), Bowden talked about their potential in replacing human calculations in governments and civil organizations (1953, p. 240), in providing solutions to commercial and business problems (1953, pp. 258–259), in predicting economic behavior and therefore the economy (1953, pp. 280–281), or in solving difficult dynamic problems in astronomy (1953, p. 317). In his opinion, "[t]he vast majority of commercially important problems can be solved perfectly well by a machine … and the same is true of many important problems in science and engineering" (1953, p. 96).

In its early history, computing seemed appealing due to its promise in improving institutional efficiency across domains, from academia to businesses to governments. At the same time, Bowden had little faith in the transformative potentials of computing machines. For instance, he doubted that computing machines could even play a game of chess well, given how badly the Manchester machine fared against a human opponent (1953, p. 93). But over the course of history, the evolution of computing was not shaped by this tunnel vision restricted to institutional efficacy. Instead, its actual historical trajectory was marked by its encounter with ideas from the cybernetics movement, and such encounter played a prominent role in branching computing research into multiple subfields.[51] To see how this turn initially came to pass, we need to situate the early development of the universal computer in the context of the cybernetics movement.

3.4. The Influence of Cybernetics on the Development of Computers

In this section, I would like to emphasize the distinction between the two intellectual traditions while clarifying that, over the history of computing, the key ideas of cybernetics did influence the direction of computer research. The distinction is important to avoid possibly confusing the universality of computing as the university of cybernetics. In the following, I will first explain the claim of universality for the Turing machine and assert that universality of the Turing machine ought not be mixed up with the interdisciplinary mosaic of the cybernetics movement. I then contend, it is due to the universality of computing that many ideas in cybernetics can be adopted in efforts to innovate the functionality of computers. Such innovations convert the original imaginary of an all-powerful calculating machine, capable of performing any computation that could be done by a human mathematician, into an intelligent machine capable of much more than mass calculation.

Prior to the cybernetics movement,[52] Alonzo Church and Alan Turing, working independently, each put forward the claim that the class of functions they defined, Church's lambda-calculus[53] and the Turing machine[54], "coincide with the informally defined class of effectively computable functions" (Cutland, 1980, p. 67). In other words, their claim, typically referred to as the Church-Turing Thesis, implies that a Turing machine, if given unlimited memory and time, "can do anything that could be described as 'rule of thumb' or 'purely mechanical'," that it can perform "every rule-of thumb

---

[51] According to Warren S. McCulloch, the Macy conferences "has been … an inspiration to neurophysiology and to the theory of automata, including artificial intelligence, and bionics or robotology" (1970, 360).

[52] This is based on McCulloch's account, according to which cybernetics "was born in 1943" (1974/2004, p. 360). There are however different opinions as to when the movement actually begins, such as the account by David A. Mindell in *Between human and machine: feedback, control, and computing before cybernetics* (2002).

[53] Alonzo Church introduced the mathematical logic of lambda-calculus in "An Unsolvable Problem of Elementary Number Theory" (1936).

[54] Alan Turing first described the Turing machine in "On Computable Numbers, with an Application to the *Entscheidungsproblem*" (1936).

process” one can conjecture (A. Turing, 1948, p. 7). In “A Note on Universal Turing Machine” from *Automata studies* (W. R. Ashby et al., 1956/1972), M. D. Davis further highlights the universality implied by the Church-Turing thesis: “it is possible to construct a definite computing machine U which is universal in the sense that any computation whatever, can be performed on U” (W. R. Ashby et al., 1956/1972, p. 167). It is worth noting that the Church-Turing thesis is only a thesis, “not a *theorem* which is susceptible to mathematical proof; it has the status of a *claim* or *belief* which must be substantiated by evidence” (Cutland, 1980, p. 67 emphasis in original). When Turing claimed that the thesis is “sufficiently well established” (A. Turing, 1948, p. 7), he was likely referring to the evidence that “[n]o one has ever found a function that would be accepted as computable in the informal sense, that does not belong to [the class of functions defined by Church and by Turing]” (Cutland, 1980, p. 67).[55]

The computation performed on a Turing machine can be another Turing machine. A universal Turing machine (UTM) is a specific type of Turing machine designed to simulate the behavior of any other Turing machine. In a modern-day computer, a software program can be considered as a Turing machine whereas the computer with its operating system can be considered a UTM. In *The Emotion Machine* (2006), Minsky gives a succinct synopsis of Turing’s idea: “[Turing] showed how to make a machine that can inspect a description of any other machine—and then interpret that description as rules for doing just what that other machine would do” (Minsky, 2006, p. 255). By switching among those different descriptions, “that same machine can, step by step, do all that those other machines can do” (Minsky, 2006, p. 256). The UTM is the predecessor of the Von Neumann architecture and other designs that are more memory and performance efficient (A. Turing, 1948, p. 7), which in turn serve as the prototypes for the design of modern-day computers.

Many critiques concerning the universality of cybernetics seem to mix up the universality of Turing machine with the loose interdisciplinary couplings over the cybernetic themes of feedbacks, homeostasis, and the blurring of boundary between the living and the machine.[56] While, as Yuk Hui (2019) has pointed out, the Turing machine

---

[55] For more on Church-Turing thesis, see Chapter 3 in Michael Sipser’s *Introduction to the theory of computation* (2013):

[56] The confusion or conflation of cybernetics and universal computing is evident in prominent works in software studies (e.g., Chun, 2013; Galloway, 2006).

has a design of recursivity similar to a cybernetic feedback loop, the universality of computing can be traced to Babbage's design of his Analytic Machine and is formally and mathematically proven by Turing. It is because of this universality that today's computing devices can "arrange our appointments, edit our texts, or help us send messages to our friends" (Minsky, 2006, p. 256). But such applications already reflect an appropriation of a computer from the original imaginary of calculating machine. We have witnessed this appropriation both in the business realm and in the personal realm. Businesses such as IBM turned the calculating machine into an information processing machine (Campbell-Kelly & Aspray, 2004). In the personal realm, it is the encounter with cybernetics that appropriates computers from performing calculations or information processing to today's personal devices. According to Thierry Bardini in *Bootstrapping* (2000), "the writings of Ashby, Wiener, and others on cybernetics deeply influenced Engelbart, then in his maturing years, just as they influenced many computer scientists in the 1950's and 1960's" (2000, p. 11). For instance, it is widely recognized that the invention of Windows and personal computing originate from Douglas Engelbart's Augmentation of the Human Intellect project. This project takes a "bootstrapping" approach of "iterative and co-adaptive learning," the basis of which "is the cybernetic notion of positive feedback in the research process" (2000, pp. 24–25). This approach adapts J. C. R. Licklider's idea of "man-computer symbiosis," which was inspired by the symbiosis of humans and machines in cybernetics (Bardini, 2000, p. 20). Bardini also notes that "[c]ybernetic concepts, methods, and metaphors gained a huge popularity" (2000, p. 11). One of the popular adaptation of cybernetics is the cyberpunk genre of science fiction, and William Gibson's vision of cyberspace in *Neuromancer* (2010) became the prominent metaphor for identifying online computer network.

3.5. Disentangling the relation between AI and Cybernetics

By tracing the early history of AI, it also seems evident that AI branched out of cybernetics. John McCarthy organized the Dartmouth summer workshop and coined the name "artificial intelligence" for this field of research to differentiate his research interest from the automata studies of the cybernetics movement. Participants invited to the workshop acknowledged the inspiration of Turing's papers on machine intelligence rather than cybernetics. Despite the differentiation, it is easy to see that the cybernetic idea of blurring human-machine boundary has left its legacy in the conceptualization of

AI. In addition, given the time of his writing, it is arguable that Turing also wrote his papers on machine intelligence under this influence of cybernetics. In the following, I will present the historical account on the Dartmouth summer workshop, Turing's reconceptualization of the UTM from a calculating machine to a machine with intelligence, and the discussion on the mechanization of the brain in the Macy Cybernetics meetings. These accounts will help us identify the similarities between AI and cybernetics as well as their differences. In particular, cybernetics emphasizes scientific understanding about biological mechanisms in the living and the artificial replication of biological mechanisms in machines. AI research, on the other hand, prioritizes the appearance of intelligence in machines, the abstraction of performing intelligent functions, over artificial replications of actual biological processes.

The Dartmouth Summer Research Project on Artificial Intelligence, which gathered together early AI pioneers in the summer of 1956, is widely recognized as the founding event of AI research ("Dartmouth Workshop," 2022). This event grew out of the reaction of John McCarthy against the work on automata theory in cybernetics. McCarthy, the initiator and organizer of the Dartmouth summer workshop, was originally one of the co-editors for the publication of *Automata Studies* (W. R. Ashby et al., 1956/1972). But he became frustrated with the type of submissions they received for the publication, as none of them resembled the exciting paths of research laid out by Alan Turing on machine intelligence (1948, 1950). It was in large part due to this dissatisfaction that McCarthy felt the necessity to organize the Dartmouth summer project, which focuses on the nouvelle and exciting works on machine intelligence distinct from the cybernetic project of automata studies. Whereas automata studies investigate the parallel between designing machines and the organizational principles of biological entities, between machines and human anatomy, the new category of research focuses on exploring how machines can imitate and exhibit human-level intelligence. Leaving behind his works with Claude Shannon on *Automata Studies*, McCarthy organized a two-month, ten-man study of artificial intelligence carried out at Dartmouth College. In the workshop, he coined the term "artificial intelligence" to distinguish this new field from automata theory, which was only concerned with "mathematical principles underlying the operation of electromechanical systems," but "not about the relation of language to intelligence, or the ability of machines to play games" (McCorduck, 2004, p. 145).

McCarthy and other participants in the Dartmouth workshop have all acknowledged the influence of Turing's two papers on machine intelligence on their research in AI. Both papers were written at around the time of the Macy Cybernetics Conferences from 1946 to 1953, and the first was written in the same year as Norbert Wiener's publication of *Cybernetics* (1948/2007), a decade or so after he conceived of the UTM. So even though the papers have no direct reference to cybernetics, it would seem highly probable that Turing was aware of the intellectual current of cybernetics and incorporated ideas from the movement when he speculated on the true potential of the UTM. In particular, his "imitation game," commonly known as the Turing Test, is likely influenced by the cybernetic theme of breaking down the rigid boundary between the human and the machine.

For Turing, the vast potentials of this universality remained largely unexplored during the early history of digital computing. Digital computers were initially deployed as immediate substitutions of human computers, as machines capable of performing mathematical calculations faster and more accurately than their human counterparts. Early applications of digital computers include the decryption of Nazi's secret communication during World War II, solving complex equations for quantum physics researchers, or the processing of numerical data in bank ledgers. But in "Intelligent Machinery" (1948), Turing "[proposed] to investigate the question as to whether it is possible for machinery to show intelligent behavior," even though its impossibility is "usually assumed without argument" (1948, p. 3). Rather than trying to "replace all the parts of [a man] by machinery," which "seems to be altogether too slow and impracticable," Turing "proposed to try and see what can be done with a 'brain' which is more or less without a body providing," directing the research focus to suitable branches of thought such as chess, poker, or translation of languages, along with cryptography and mathematics (1948, p. 13). In "Machine Intelligence" (1950), Turing further problematizes the question on intelligence by reframing the question in terms of what he calls the "imitation game" (1950, p. 1), or what people today call the Turing Test. This hypothetical game involves three players: an interrogator, a human, and a digital computer. The interrogator is in one room, the person and the computer in another room. The interrogator can pose questions to either the human or the computer, and would win the game by correctly discerning which is the human and which is the computer. But the game is not so easy because the computer would simulate human behavior to thwart attempts by the interrogator to discern between the two, hence the name "imitation game." This game illustrates how Turing wants to redefine intelligence: A machine can be said to possess human-level intelligence as long as people cannot tell how its behavior is different from human behavior, regardless of the causes that lead to such behavior.[57]

The cybernetic counterpart to destabilizing the meaning of intelligence is the idea that human brains are biological machines and can be substituted by artificial brains, not unlike how our heart can be replaced, at least temporarily, by an artificial heart. Such a controversial idea was a major topic of debate in the Macy conferences. For instance, Ralph W. Gerard asserts, "to say, as the public press says, that therefore these machines are brains, and that our brains are nothing but calculating machines, is presumptuous" (Gerard, 1950/2003, p. 172). Donald M. MacKay speculated the possibility of simulating consciousness with a probabilistic mechanism (1951/2003). In his presentation of information theory, MacKay suggested that "[c]onsciousness … —if I dare stick my neck out—might be introduced in this way: We might say that the point of area »of conscious attention« in the field of view—in a field of data—is the point or area under active symbolic replication, or evocative of (internal) response" (1951/2003, p. 494). He believed that "one could go a very long way toward simulating what appears to be the ordinary conscious behavior of human beings" by "[devising] a probabilistic mechanism with the same mobility, and so on, as Homo sapiens, if one would have to go in for mechanisms in protoplasm instead of mechanisms in copper" (1951/2003, p. 495). John von Neumann continues this line of thought in *Autonomous Studies* (W. R. Ashby et al., 1956/1972), where he talked about the possibility of a UTM mimicking the human intuition: "[I]f our automata are furnished with an unlimited memory — for example, an infinite tape, and scanners connected to afferent organs, along with suitable efferent organs to perform motor operations and/or print on the tape - the logic of constructable machines becomes precisely equivalent to intuitionlistic logic" (1956/1972, p. 50).

---

[57] Over the years, the Turing test has been widely challenged and criticized. For instance, John Searle's Chinese room argument implements a version of the Turing Test to show that the test is insufficient to detect the presence of consciousness, even if the room can behave or function like a conscious mind. Here, I will not dwell on the philosophical debate on whether or not a machine can truly think like a human being. I simply want to highlight how influential Turing's thesis had been for the AI pioneers who met at the Dartmouth Summer workshop.

Between the conference participants, there seems to be a tension between those who want to build electronic brains or artificial organs, and those who want to build machines as models of living organisms in order to study their scientific properties. It is a tension between an engineer's aspiration and a scientist's quest for knowledge. According to "A Note by the Editors" for the Eighth Cybernetics Conference, the eventual consensus of the conference participants was to prioritize scientific knowledge over functioning machines:

> We [participants of the Macy conferences] all know that we ought to study the organism, and not the computers, if we wish to understand the organism. … But the computing robot provides us with analogues that are helpful as far as they seem to hold, and no less helpful whenever they break down. To find out in what ways a nervous system (or a social group) differs from our man-made analogues requires experiment. These experiments would not have been considered if the analogue had not been proposed, and new observations on biological and social systems result from an empirical demonstration of the shortcomings of our models. (von Foerster et al., 1951/2003, pp. 346–347)

Along the same line, Pitts suggested that the purpose of constructing machines is to model after mechanisms of biological organs or of the nervous system, and these models serve two functions:

> First they want to demonstrate that thus and such a function … can in fact be done by some mechanism. This is an extremely important educational task, since certainly the vast majority of the world would refuse to accept for a moment the assumption that what they regard as specifically psychological functions can be done by any mechanism whatsoever; … The second function of modelmakers is to find models that throw a light, either directly in the sense of making a mechanism out of components as much like what we know about the neurons as possible, so that we can perhaps form direct suggestions as to how in fact the brain does something, or else indirectly, by means of elements which have certain of the formal properties … (Gerard, 1950/2003, p. 649)

The first function is to demonstrate that certain psychological functions can be achieved through mechanisms. This function is important because, to many people, psychological functions are unique to humans and cannot be replicated by machines. The second function is to shed light on how the brain functions, either by directly mimicking the structure of neurons or by using elements that have similar properties.

It is in this sense that cybernetics research and automata studies differ from research on artificial intelligence. The primary goal of cybernetics is to discover the mechanisms behind the "human machine." The possibility to construct machines that replicate the "human machine" was regarded as both a consequence of and an instrument for this discovery. It assumes a symmetric view of the mechanisms between the living and the non-living. In contrast, AI is a field that favours simulation of (human) behavior over the imitation of mechanisms. It values an algorithm that generates an appearance of intelligence even if the algorithm has no matching biological or physiological processes.[58] This emphasis on abstraction and encapsulation comes from computer sciences, as the two concepts are fundamental design principles in hardware and software architecture.[59] In Chapter 4, we will further explore this distinction between the cybernetic machine that learns and the abstract machine-learning models in AI.

3.6. Summary

In this chapter, I question the critique of universality of cybernetics by distinguishing between a macro-social perspective that attribute the emergence of our technocratic society to the cybernetics movement, and a technical perspective that recognizes the irresolvable conflicts between disciplines loosely tied by cybernetic concepts like feedbacks or homeostasis. I first problematize the notion of universality in cybernetics by reading the actual conversations that took place at the Macy Cybernetics Conferences. The conversations reveal irresolvable conflicts and tensions between conference participants, mainly due to the inherent conflicts in the presumptions held by various disciplines. The interdisciplinary meetings were loosely held together by a few key themes, such as feedbacks or homeostasis, which could be interpreted in a variety of ways between the conference participants. The loose ties would not have been

---

[58] This view on the relationship between cybernetics and AI is coherent with Roberto Cordeschi's synopsis in *The Discovery of the Artificial* (2011): "The fundamental insight of cybernetics, i.e. the proposal of a unified study of organisms and machines, was inherited, starting in the mid-1950s, by AI (Artificial Intelligence). However, AI proposed a different simulative methodology. To put it quite generally, this methodology used computer programs to reproduce performances which, if observed in human beings, would be regarded as intelligent" (2011, p. xi).

[59] In *The Closed World* (1997), Paul N. Edwards makes a similar point: "Instead of modeling brains in computer hardware—the central goal of cybernetics—AI sought to mimic minds in software. This move from biological to symbolic models has usually been interpreted as an abrupt intellectual break, a sudden shift in orientation from process to function" (1997, p. 239). But Edwards' emphasis is placed on the contrast between the physical (brain, biological, hardware) and the virtual (mind, symbolic, software), whereas my account emphasizes the cyberneticists' goal in finding parallels between the biological and the machine (both software and hardware), in contrast to researchers in artificial intelligence who see little value in such parallels.

possible without a scientific discourse that already assumes a mechanistic worldview. Scientists shared the exploratory aim of trying to discover the mechanisms behind animal, human, and social behaviour.

I then trace the history of computing and of the UTM to Babbage's Analytic Machine, which was originally designed to replace the mass calculation by a factory of human computers. As we can see in Bowden's *Faster than Thought* (1953), this perception of the computer as a super calculator was the prevalent technological imaginary during the 1950s. But already, the intellectual current of cybernetics is already steering the research direction of computing to other directions, among them, the human-computer symbiosis in Licklider's and Engelbart's research that led to the invention of windows and personal computing, as well as the research on artificial intelligence.

Particularly relevant to this dissertation is the relation between cybernetics and artificial intelligence. We can see traces of cybernetics themes in Turing's papers on machine intelligence, in which he proposes the "imitation test" to destabilize the meaning of intelligence, traditionally regarded as a unique human quality. The transactions of the Macy conferences also reveal a similar challenge to the traditional boundary between the human and the machine, discussing the mechanistic understanding of the human brain and the possibility of replicating these mechanisms in a machine. Nonetheless, cybernetics emphasizes scientific understanding that reveals actual biological mechanisms over functionable technological designs. Conversely, AI research is only concerned with the simulation of human behaviour, and values abstract algorithms even if they have no resemblance to the mechanisms that sciences reveal in the human body or the nervous system. This reflects the design principles of abstraction and encapsulation commonly found in computer hardware or software architecture.

Machine learning, a subfield of AI, also bears this complex relationship to cybernetics. In fact, machine learning has closer ties to cybernetics than the rest of the discipline in AI and in computer sciences. Machine learning adopts the semantic change of the word "learning" in cybernetics, and there are strong resemblances between machine learning in AI and the cybernetic learning machines, such as Claude Shannon's maze-solving electromechanical mouse or Ross Ashby's Homeostat. Nonetheless, machine learning in AI does not rely on scientific knowledge of biological or physiological mechanism, but rather, explores and attempts to find the best abstract algorithms that can be programmed on a UTM. It eventually evolves into a field of knowledge about algorithms and statistical/probabilistic proofs on predictive models based on pattern recognition. I will further elaborate on this genealogy of machine learning in the following chapter.

Chapter 4.

The Genealogy of Machine Learning: From Learning in Cybernetics to Machine Learning in AI

Me: is there any relationship between machine learning and cybernetics?

ChatGPT: Yes, there is a relationship between machine learning and cybernetics. Cybernetics is the study of control and communication in systems, whether they are mechanical, biological, or social. Machine learning, on the other hand, is a subfield of artificial intelligence that involves building algorithms and models that can learn and make predictions or decisions based on data.

Both machine learning and cybernetics are concerned with systems that can adapt and learn from their environment. Cybernetics provides a theoretical framework for understanding and designing systems that can control themselves, while machine learning provides practical tools for building these systems. In particular, machine learning algorithms can be used to design control systems that can adapt and improve based on feedback from their environment.

4.1. Overview

Back in Chapter 3, I contended that the emergence of AI comes from the encounter of two intellectual traditions. One is associated with the universal computing machine that was originally designed to replace the factory of human computers who collectively compute arithmetic. The other is associated with the themes of feedback and boundary-crossing of the cybernetics movement. One of the major topics discussed in the Dartmouth summer workshop on AI is machine learning. This research direction, led by Arthur Samuel and Ray Solomonoff, actually inherits more directly from cybernetics than the rest of the discipline in AI and in computer sciences. Prior to the Dartmouth meeting, cyberneticians had been exploring whether and how a machine can "learn," as evident in the transactions of the Macy conferences and in the publication of *Automata Studies* (1956/1972). Turing's papers on machine intelligence, which also present a lengthy discussion on how a universal Turing machine can be made to "learn," was arguably written under the influence of the cybernetics movement, In fact, the very name "machine learning" seems to imply this cybernetic heritage.

At the same time, machine learning, as a subfield of AI, is distinguishable from the learning in cybernetic machines and in automata studies. Research on machine learning over the past half-century has primarily developed software-based algorithms, unlike the electromechanical designs of cybernetic learning devices such as Claude Shannon's maze-solving mechanical mouse or W. Ross Ashby's Homeostat. These devices were designed to replicate the learning mechanisms in living organisms. In contrast, machine-learning rides on the mathematical properties of the universal Turing machine. It is by realizing the implication of this universality that Solomonoff came up with the proofs on algorithmic probability and universal inductive inference. These proofs confirm that machine learning is capable of recognizing any subtle patterns that exist in a sea of data, establishing that the primary affordance of machine learning is pattern recognition.

We will come back to Solomonoff's proofs in Chapter 5. This chapter attempts to elucidate the legacy of learning machines in cybernetics on the AI subfield of machine learning as well as the way machine learning departs from cybernetics. Going back to the transactions of the Macy Cybernetics Conferences, I will show how the discussion on Norbert Wiener's guided missiles and statistical predictors, Shannon's maze-solving device and that on Ashby's Homeostat engage in the negotiation of what "learning" can mean in the context of machines. The resulting semantic change in "learning" was adopted by the AI subfield of machine learning, which has also developed a statistical framework of training models based on cybernetic feedbacks and homeostasis. Nevertheless, just like symbolic AI, which was the dominant approach to implementing AI in its early days, machine learning is concerned with formulating abstract knowledge on computer algorithms and mathematical proofs about these algorithms. Unlike cybernetics, these algorithms do not necessarily imitate the mechanistic knowledge of the actual human body or the nervous system. Such an emphasis on abstract algorithms is evident in Turing's papers on machine intelligence, in Samuel's checkers-playing machine, and in Solomonoff's theories.

4.2. The Mechanistic “Learning” of Cybernetic Feedbacks

The idea of building machines that can “learn” by adapting to environmental feedbacks was widely discussed in the Macy Cybernetics Conferences. Participants across different scientific disciplines would compare the type of learning observed in humans or in animals with the mechanistic “learning” in cybernetic systems, which can be achieved by implementing feedback loops in a system design. Such systems would give the appearance of “learning” when their internal parameters are tuned in response to feedbacks from their environment. By generalizing “learning” as a feedback phenomenon for both the living and the machine, and by redefining “adaptation” as the characteristic of a system capable of self-reorganization based on feedbacks from its environment, the cyberneticians were implicitly suggesting semantic changes in the words “learning” and “adaptation,” further problematizing the boundary between the living and the machine. In particular, the systems invented by three cyberneticians show off this twist in the semantic of “learning” and became the focal points of discussion during the Macy conferences. In the following, I will first depict Norbert Wiener’s invention of guided missiles and his subsequent project of statistical-based predictors. These systems are prototypes of artificial “learning” based on negative feedbacks. I then present Claude Shannons’s maze-solving electromechanical “mouse” and W. Ross Ashby’s Homeostat to illustrate how other cyberneticians have adopted Wiener’s method of negative feedbacks in their learning and adaptation machines.

The design framework of artificial “learning” typically involves a feedback loop between a device’s input, output, and memory storage. In fact, the first archetype of a feedback system, Wiener’s guided missile, can also be perceived as a learning system. During World War II, Wiener was designing missiles that could be guided into the future position of an enemy airplane. These “goal-seeking” missiles “predict the future position of a moving target (at time of impact) by extrapolation from its earlier positions during pursuit” (von Foerster et al., 1951/2003, p. 345). Extrapolation refers to how “the missile measures the angle between its direction and the target it’s seeking” (Brand, 1976/2004, p. 302). The missile uses this measurement to continuously correct its path. Making predictions in a targeted manner based on feedbacks of behavioral patterns is in fact the hallmark of Wiener’s cybernetics as exhibited by his guided missiles. The missile is learning, in real-time, the pattern of how its target moves, and by recognizing this pattern, it can make reasonable predictions about the future positions of the target.

Wiener is a mathematician, and his research has primarily been statistical. Such works have led to his insight regarding the possibility of Big Data. At the Second Cybernetics Conference in 1946, Wiener explained the intuition that more data would always yield better predictions. Without going into rigorous mathematical proofs, he stated that “from any time series it was possible to compute a best prediction, that the more data were available the better became a prediction” (McCulloch, 1947/2004, p. 341). From this intuition, Wiener deduced that “wherever sociologists or social anthropologists were able to collect time series which need be no more than enumerations of decisions at specified times, it would be possible to discover causal sequences in human conduct” (McCulloch, 1947/2004, p. 341). This theoretical possibility of Big Data, however, was deemed impossible in practice at the time due to lack of data and computational resources.[60]

Wiener’s statistical works are also precursors to the statistical framework in machine learning. At the Sixth Cybernetics Conference held in 1949, Wiener declared that he had been preoccupied with building a statistical predictor “which will actually examine its own statistical experience and do its circuit in accordance with the statistical experience if the data changes in character and will change itself to suit the new statistics and data” (Pias, 1949/2003, p. 158). This predictor would “go through the computing motion, compute the auto-correlation and the entire set of patterns for prediction theory … continually changing its pattern of prediction if there is any change in the statistical pattern of the data” (Stroud, 1949/2003, p. 43). The machine would compare predicted error with the actual error, and “when the difference between the actual error and the predicted error got beyond a certain percentage—it would repeat itself, re-examine itself statistically” (Stroud, 1949/2003, p. 44). By iteratively measuring the accuracy of outputs and then reconfiguring parameters, Wiener’s statistical predictor employs an algorithm similar to today’s machine-learning algorithms. In a typical

[60] In Warren S. McCulloch’s summary of the beginning of cybernetics, he presents Wiener’s view that, even though there are cybernetic problems in the study of social behavior, we “lack sufficiently long runs of uncorrupted data to apply mathematical tools” (McCulloch, 1974/2004, p. 359). Today, insufficient data is no longer an issue in a society where a large portion of social interactions occurs on the surveillance Internet. Wiener’s improbable idea of discovering causal relations in social behavior has now become a practical reality.

machine learning technique, cost functions are computed to model learning problems, and optimization schemes such as gradient descent would iteratively tune models' parameters with the goal of minimizing the cost.

In "A Note by the Editors" for the Eighth Cybernetics Conference, Heinz von Foerster et al. note that "the appearance of 'purpose' in [complex electronic devices'] behaviour (a feedback over the target!) has intrigued the theorists and prompted the construction of such likeable robots as Shannon's electronic rat …" (1951/2003, pp. 345–346). Claude Shannon is renowned for his theoretical contributions to information theory. But he also played a significant role in the research on automata, which consist of both theoretical and engineering works for building automated machines. One such machine is the maze-solving electromechanical mouse called "Theseus." Built in 1950, Theseus is probably the first artificial learning device of its kind (*Bell Labs Advances Intelligent Networks*, 2012). It is a mechanical mouse controlled by an electronic relay circuit that controls driving motors through a feedback loop with its sensing finger (Shannon, 1951/2003, p. 477). It can move around a labyrinth of 25 squares with flexible maze configurations, which can be modified by rearranging movable partitions. Theseus can be placed anywhere and, by trial-and-error, finds its ways toward the goal. If it has travelled through the maze, the electromechanical circuit would retain the "memory" of the maze configuration, allowing it to go directly toward the goal next time around. If Theseus is placed in a square that it has not explored, or if partitions have been reconfigured, it would search its way by trial-and-error to a known location and then proceed directly toward the goal.

Shannon presented his electromechanical mouse at the Eighth Cybernetics Conference because he believed that other conference participants might be interested in "its connection with the problems of trial-and-error learning, forgetting and feedback systems" (Shannon, 1951/2003, p. 474). Theseus was designed with a large number of feedback loops, the most prominent one being the feedback "from the sensing finger through the circuit to the driving motors and back to the sensing finger, by electromechanical motion of the motors" (1951/2003, p. 477). Its ability to learn is captured in another feedback loop, one between the electromechanical mouse's memory, its driving motor, and its sensing finger. The device "is capable of remembering one of four possible directions: north, east, south, or west" for each square in the maze

(1951/2003, p. 476). The feedback of this memory into the device's strategy to directly reach the goal is what makes Theseus an artificial learning device.

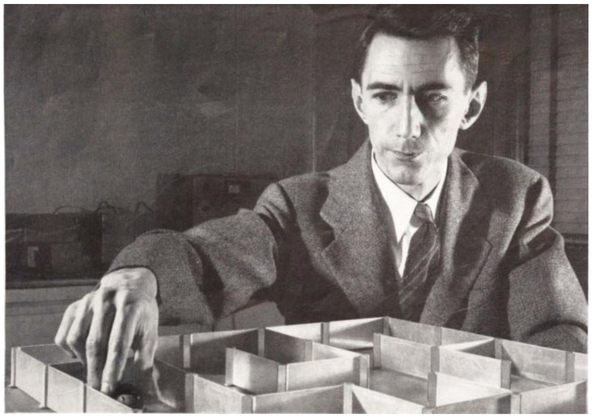


Figure 1: Claude Shannon and its electromechanical mouse (Bell Lab, n.d.).

It is commonly acknowledged that "[n]egative feedback is a central homeostatic and cybernetic concept, referring to how an organism or system automatically opposes any change imposed upon it" (Rodolfo, 2000). The homeostat, invented by W. Ross Ashby, is another cybernetic system designed to learn from feedbacks. It is a mechanical apparatus that could regain stability entirely on its own in adaptation to a dynamically changing environment. It models after the biological notion of homeostasis in living organisms and was designed for a single purpose: When disturbances are introduced into the system, the device would automatically reconfigure itself to stabilize the effects of the disturbances. Cyberneticians adopted this biological notion and converted it into a technological notion such that homeostasis can be modelled in machines.

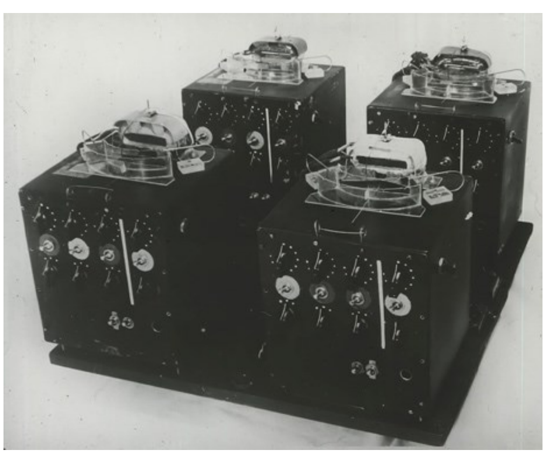


Figure 2: W. Ross Ashby's Homeostat (M. A. Ashby on behalf of the Estate of W. Ross, 1948)

Ashby completed the construction of the homeostat in 1948, in the same year when the intellectual current of cybernetics was christened by Wiener with his publication of *Cybernetics* (1948/2007). Wiener's book gives a brief overview on the mechanistic understanding of homeostatic processes by alluding to the feedback systems of living organisms (see 1948/2007, pp. 114–115). He emphasizes the importance of homeostasis in cybernetics, suggesting that "[a]ny complete textbook on cybernetics should contain a thorough detailed discussion of homeostatic processes" (1948/2007, p. 115). But as an introductory survey into cybernetics, his book has only provided "an introduction to the subject [rather] than a compendious treatise" (1948/2007, p. 115). The task of writing this "compendious treatise" fell on Ashby, who explains the cybernetic notion of homeostasis and describes his design of the homeostat in the article "Design for a Brain" (1948) and later in a book under the same title (1952/1960). In the same year his book was published, he gave a demo of the homeostat at the Nineth Cybernetics Conference. In the demo, he explained the difference between the feedbacks in Wiener's guided missiles and those in the homeostat. Even though the homeostat and Wiener's guided missiles are both feedback systems, they can be distinguished in the way the system changes: "During learning, the organism must change, but this change is not to be confused with the change that it undergoes during its small corrective movements" (W. R. Ashby, 1952/2003a, p. 610). Instead of undergoing small corrective movements like Wiener's guided missiles, the homeostat imitates the self-regulating process of homeostasis, the way an organism adapts to its environment by reorganizing its internal structure. Thus Ashby "consider[s] the organism … as a mechanism which faces a hostile and difficult world and has as its fundamental task keeping itself alive" (1952/2003a, p. 593). He was interested in discovering the mechanisms that can bring an equilibrium between internal organization and external environment for "keeping itself alive," and considered the environment as "a transducer, as an operator that converts whatever action comes from the organism into some effect that goes back to the organism" (1952/2003a, p. 594). Because an organism needs to survive by "developing an adaptive reaction to any one of an almost unlimited number of environments" (1952/2003a, p. 595), the organism needs to have an internal structure that is most flexible and adaptable to environmental changes. It needs to switch between internal structures when faced with a new environment. Thus "[t]he fundamental problem [of learning] is one of organization, of finding the appropriate switching-pattern" (1952/2003a, p. 595). Herhardt von Bonin identifies the same adaptive self-reorganization in animals' behaviour: "the animal appears to break down the environment into certain patterns which seem to develop in its mind. Something goes on in its brain and then that structures the environment. It perceives the environment in a certain pattern, set by its brain, so that it can deal with it" (1952/2003a, p. 603).

For Ashby, the significance of the homeostat is not limited to proving the feasibility of building a cybernetic machine with homeostatic processes. The homeostat can actually serve as a prototype for an artificial brain. In the article "Design for a Brain" (1948), he speculated that a perfected homeostat is capable of playing chess "with a subtlety and depth of strategy beyond that of the man who designed it." *Time* magazine concurs, describing the homeostat as "the closest thing to a synthetic brain so far designed by man" ("Science: The Thinking Machine," 1949). A few years later, Ashby developed this idea more comprehensively in the book *Design for a Brain* (1952/1960). In the preface of the book, Ashby gives the reasons behind why the homeostat can be considered a model for the brain:

> The book is not a treatise on all cerebral mechanisms but a proposed solution of a specific problem: the origin of the nervous system's unique ability to produce adaptive behaviour. The work has as basis the fact that the nervous system behaves adaptively and the hypothesis that it is essentially mechanistic; it proceeds on the assumption that these two data are not irreconcilable. It attempts to deduce from the observed facts what sort of a mechanism it must be that behaves so differently from any machine made so far. (1952/1960, p. v)

Among the different cerebral mechanisms, Ashby focuses on the mechanism that is unique to the nervous system: its ability to produce adaptive behaviour. His goal is to discover the mechanism behind this adaptive behaviour, and the discovery would allow him to build a machine that is unlike any machine before, a machine that is capable of behaving like the nervous system and the human brain. Thus the homeostat was built to replicate and to demonstrate the mechanisms behind the cerebral mechanisms that produce adaptive behaviour in living organisms (1952/1960, p. 99).

In another meeting at the Ninth Cybernetics Conference, Ashby presented more insights on the homeostatic processes of cerebral mechanisms in his presentation on "Mechanical Chess Player" (1952/2003b). Without having constructed an actual device, Ashby presented his homeostatic design of a mechanical chess player that "learns" from its environment based on pattern recognition. He characterized "learning" as the ability to “develop better criteria of judgement than the designer himself can produce” (W. R. Ashby, 1952/2003b, p. 651). This ability can be developed through “corrective feedback that is operated by results” (W. R. Ashby, 1952/2003b, p. 652). Ashby applied this concept of learning automaton to the design of a mechanical chess-player. This design has less to do with typical search algorithms (such as “minimax algorithm” or “alpha-beta pruning”) than with the recognition of layout patterns of chess pieces on a chess board. The idea occurred to Ashby when he came upon *Capablanca’s Chess Fundamentals*, where he “found many sentences each of which gave clear advice in a general way without making any specific analysis on specific squares” (W. R. Ashby, 1952/2003b, p. 653). This made Ashby realize that the growth in recognizing patterns on a chess board is how a beginner develops into an experienced chess player. In analogy, a mechanical chess player can be a homeostatic system that continually improves its strategies based on corrective feedbacks with its environment, which is simply the layout patterns of the chess pieces on a chess board. Thus Ashby’s conceptual design of a mechanical chess player exemplifies an evolving homeostatic system that adapts and learns by recognizing patterns in its environment.

4.3. The Controversy of Semantic Change in “Learning”

From Wiener’s guided missiles to Shannon’s mechanical mouse to Ashby’s homeostatic devices, these feedback-based systems can dynamically reach certain goals by modifying their internal states. During the Cybernetics meetings, participants deliberated on whether the appearance of purposes in artificial systems can be considered as a form of learning. The semantic maneuver would subvert the traditional understanding of learning as a capability that belongs solely to the living. While the conference participants failed to reach any consensus on the matter, the semantic change in the word “learning” is taken up later when the term “machine learning” was coined for the field of study concerned with the training of computers with data to improve their performance over time without being explicitly programmed (Samuel, 1959). The coining of "machine learning" would presumably help differentiate this field from other areas of computer science and AI.

In the Cybernetics meetings, scientists and scholars from different disciplines shared their disciplinary perspective on various forms of “learning.” For instance, the developmental psychologist Herbert Birch was interested in

> [w]hat kind of learning is necessary for an organism to generalize, to abstract from concrete experience, certain gestural relations which are relevant to these experiences? So we have, then, the problem of the study of the evolution of intelligence in organisms; that is to say, the study of the development of modifiability. (Birch, 1951/2003, p. 464)

Learning in this biological context is less associated with goal-directed predictions and more associated with the phylogenetic or ontogenetic evolution of intelligence in organisms (see Birch, 1951/2003, pp. 463, 471). Phylogenetic evolution is concerned with the improved adaptability of a species to its environment across generations. Ontogenetic evolution is concerned with the growth of an individual through its experience. Both kinds of evolutions involve feedback interactions between the organism or species with its environment. In order to survive, it must maintain a relatively stable equilibrium over these feedback loops.

Maintaining stable equilibrium involves the notion of homeostasis, but it is debatable whether adaptations alone would necessarily lead to either ontogenetic or phylogenetic evolution, or any other form of learning. During the conferences, Ashby often got into arguments with other participants on whether the adaptability of the homeostat ought to be classified as a genuine form of learning. Julian Bigelow, who co-authored with Wiener and Arturo Rosenblueth one of the founding papers on cybernetics, "Behavior, Purpose and Teleology" (Rosenblueth et al., 1943), was one of those uncomfortable with the mechanistic notion of learning. He challenged Ashby, “[t]he machine [homeostat] finds a solution, I grant you … I merely wonder why finding a solution necessarily implies that it learns anything” (W. R. Ashby, 1952/2003a, p. 615). This led to a heated debate on the semantic modification of the word “learning” (see W. R. Ashby, 1952/2003a, pp. 615–616). J. Z. Young expressed that “[t]he essence of learning is that the system that has been through a procedure has different properties than those it had before” (W. R. Ashby, 1952/2003a, p. 615). Jerome B. Wiesner felt that a mechanism is learning “if on the next trial the searching is not completely random” (W. R. Ashby, 1952/2003a, p. 616). After listening to these attempts to qualify learning mechanistically, Bigelow still found it “difficult here to associate this way of finding a solution with the word learning” (W. R. Ashby, 1952/2003a, p. 616).

As with many topics discussed in the Cybernetics conferences, participants were unable to arrive at a consensus in this debate. What we are witnessing, though, is the historical process of destabilizing the traditional meaning of “learning” in the proposal of a mechanistic notion of learning.[61] This mechanistic notion was later adopted by the particular branch of computer science and AI called machine learning. Machine learning today, as a subfield of AI, has adopted the semantic changes of “learning” that patterns after the ontogenetic or phylogenetic evolution of intelligence in organisms. Many systems embedded with machine-learning component would continually adapt and reconfigure internal parameters based on new input signals from their environments, in the same way as Ashby’s homeostat adapts to its environment by recognizing data patterns. The legacy of cybernetics on machine learning is undeniable. At the same time, along with other areas in AI research, machine learning departs from cybernetics in its methodology. Samuel introduced the field as the studies “concerned with the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning” (1959). As I will elaborate next, works in machine learning explore the potentials of universal computing to implement mechanistic learning. These works typically come up with abstract algorithms that are unrelated to actual biological or physiological processes in living beings.

4.4. Differentiating Machine Learning from Learning in Cybernetics

As I remarked back in Chapter 3, AI emerged from the encounter of universal computers with the intellectual current of cybernetics. Negotiating a new connotation of “learning” in machines is among the legacies left behind by cybernetics, taken up by AI research in the machine-learning subfield. Between learning in cybernetic machines and machine learning in AI, there is a subtle difference in their respective focus. Cybernetics devotes efforts in designing and building learning machines that are similar to living organisms. It upholds an approach that treats the living and the non-living symmetrically. In contrast, AI is a field that does not favour imitation over abstract simulation. It values

[61] In *On the Origins of Cognitive Science: The Mechanization of the Mind* (2000), Jean-Pierre Dupuy argues that cybernetics represents not the anthropomorphization of the machine but the mechanization of the human. Accordingly, the early cyberneticians intended to construct a materialist and mechanistic science of mental behavior, and this construction was influential to the birth of cognitive science. At the time, cognitivism was a movement in psychology in response to behaviorism, which identifies thinking as a behavior and neglects to explain cognition. This shift from behaviorism to cognitivism has affected the design of computer-assisted learning (CAL) (Hartley, 2006).

an algorithm that generates an appearance of intelligence even if the algorithm has no matching biological or physiological processes. As a subfield of AI, machine learning sets their vision on creating a computing agent capable of human-like intelligence. It has evolved from the imitation of physiological knowledge of nervous systems to the abstract knowledge of machine-learning algorithms (e.g., gradient descent, support vector machine, kernel method, or the mathematical properties of the neural network[62]). The ability of a machine to "learn" from data is both a necessary trait of a super-intelligent computer and a possible path for creating computers that can attain super intelligence.

Since its inception at the Dartmouth summer workshop, AI research has turned away from the cybernetic approach, in which the human body and its nervous system would be investigated under the lens of mechanisms that can be replicated in a machine construction. Instead, AI research explores the potential of the universal computer in simulating human intelligence. The dominant approach in its early days was symbolic AI, which is "based on high-level symbolic (human-readable) representations of problems, logic and search" ("Symbolic Artificial Intelligence," 2022). According to Grace Solomonoff, the wife of Ray Solomonoff, the Dartmouth summer workshop was at least partially responsible for the turn toward symbolic AI (2019, p. 16). Over the course of the Dartmouth summer workshop, Ray Solomonoff convinced Marvin Minsky to turn his research focus from the mechanical imitation of the brain (the neural network) to symbolic representation (McCorduck, 2004, p. 101; G. Solomonoff, 2019, p. 23). With their goal of building machines that can appear intelligent, it would seem much more feasible to achieve this goal with symbolic representations programmed in software. Symbolic AI became the de-facto method for AI research after Herbert A. Simon, J. C. Shaw, and Allen Newell demonstrated their General Problem Solver in 1957. The General Problem Solver is a computer program designed to solve problems in a general way rather than being limited to a specific domain. It can find solutions to mathematical problems such as algebraic equations, solve puzzles such as Sudoku puzzles, or generate plans for complex tasks such as assembling a machine. With its tangible success, General Problem Solver sparked much popular enthusiasm in AI during the 1960s and 1970s. Since then, symbolic AI has been used to design and build

[62] I will further explain how the neural network has transitioned from a modeling of a body's nervous system to an abstract model for mathematical proofs in Chapter 5 when I discuss the implication of Ray Solomonoff's Algorithmic Probability.

“knowledge-based systems (in particular, expert systems), symbolic mathematics, automated theorem provers, ontologies, the semantic web, and automated planning and scheduling systems” (“Symbolic Artificial Intelligence,” 2022).

In comparison with symbolic AI, machine learning brings a different approach to developing machine intelligence. It differs from symbolic AI in the conviction that the machines can be “programmed” by data better than by programming-by-hand, and that in the long run, machine learning can attain a higher level of machine intelligence than symbolic AI. This scheme of machines “learning” from data, along with the statistical and probabilistic nature of machine-learning, are all inherited from cybernetics research. Comparing to symbolic AI, cybernetics has been much more influential to the technical knowledge developed for machine learning. At the same time, like symbolic AI, machine learning departs from the cybernetic approach, exploring the potential of universal computers rather than designing machines that replicate biological or physiological processes. Thus most machine-learning techniques can be distinguished from the mechanistic replications of such processes in cybernetic machines.

Already in Turing’s “Intelligent Machinery” (1948), we can see the seed of computer-based machine learning taking shape in his proposed design of software programs capable of “learning.” In the first part of this paper, Turing deliberates on how to build intelligent machines “designed for a definite purpose” (1948, p. 113), which is basically the approach of symbolic AI. Halfway through his paper, Turing begins to describe an alternate approach, the construction of “unorganized machines,” which are “largely random in their construction” as they are made up “in a comparatively unsystematic way from some kind of standard components” (1948, p. 9). In Section 1.4, I explain how Dreyfus’ critique was primarily targeted at the formalist approach symbolic AI and advocated a research direction on “the forms of ‘information processing’ essential in dealing with our nonformal world” (Dreyfus, 1992, p. 216). Apparently, Dreyfus was not aware of the non-formalist, unsystematic approach that Turing already outlined in the second half of his seminal paper on machine intelligence. This “comparative unsystematic” approach anticipates later works in machine learning, which has today proven to be a form of information processing capable of dealing with much of our nonformal world. It leaves open the possibility of interference, which if applied appropriately can “[mimick] education” of a learning machine (1948, p. 14). This design is analogous to “the cortex of the infant,” which “is an unorganized machine, which can be organized by suitable interfering training" (1948, p. 16). One can therefore "experiment with unorganized machines admitting definite types of interference and try to organize them, e.g., to modify them into universal machines" (1948, p. 17). Turing then suggests, just as "the training of the human child depends largely on a system of rewards and punishments … it ought to be possible to carry through the organizing with only two interfering inputs, one for 'pleasure' or 'reward' (R) and the other for 'pain' or 'punishment' (P)" (1948, p. 17). This scheme is now known as supervised learning in machine learning. The "organization of a machine into a universal machine" (1948, p. 17), in today's technical language, would be the training of a machine-learning model into a specific computer program with well-defined behavior.

In contrast to cybernetics research, bringing up child-like learning in a discourse of machine intelligence is metaphorical for Turing. This metaphorical discourse is reiterated in "Computer Machinery and Intelligence" (1950). Framing the discussion in the context of the imitation game, Turing asks whether the machine imitating the human can be a program simulating the child's mind, taught through an education process, rather than a program simulating the adult's mind (1950, p. 20). In this paradigm, Turing accurately predicts an intrinsic opacity of reasoning in machine learning, a characteristic that many social critics today find most concerning:[63] "Most of the programmes which we can put into the machine will result in its doing something that we cannot make sense of at all … Intelligent behavior presumably consists in a departure from the completely disciplined behavior involved in computation" (1950, p. 21).[64]

Arguably, early researchers on machine learning took the lead from Turing's papers and continued developing his ideas in their research. One of them was Arthur Samuel, a key participant in the Dartmouth Summer Project. He completed his construction of a checker-playing machine that is commonly recognized as the first successful application of machine learning. This program could improve its gameplay after many rounds of playing to eventually reach the level of an average human player. Minsky remarked that Samuel's checkers-playing program was "at present the outstanding example of a game-playing program that matches average human activity"

---

[63] See Berghoff et al., 2021, pp. 17–18; Brown et al., 2021, p. 5; Carabantes, 2020; Chan, 2021; Fainman, 2019; E. Lee et al., 2021; Müller, 2021; Watson, 2021; Zednik & Boelsen, 2021.

[64] This opacity of machine reasoning is exemplified in Ray Solomonoff's algorithmic probability, which we will further discuss in Chapter 5.

(1961). Samuel's original proposal in 1947 was to build a good, "intelligent" checkers-playing machine, but not a self-learning machine (McCorduck, 2004, pp. 173–180). He was exploring different optimization heuristics[65] to achieve this goal, and the heuristics that worked out best was the "success-reinforcement decision model," which allowed the machine to improve its checkers-playing ability by simply playing against itself to the point of rivaling average human ability, without relying on pre-programmed strategies.[66]

Another pioneer of machine learning is Ray Solomonoff, whose works have been theoretical throughout his life, devoting his entire career to the computational theories and mathematical proofs on machine-learning models. Unlike Samuel, he was less interested in optimizing heuristics for building an actual functioning machine than in exploring a general theory for machine learning. He developed mathematical proofs on the potentials and the limits of what an "unorganized machine" on a universal computer can learn. For many years, machine learning has subsisted as a niche, marginal subfield of AI research due to the lack of real-world applications for businesses and consumers. While Dreyfus quickly dismissed Solomonoff's mathematical theory of inductive inference as a desperate and futile attempt at the self-improvement and learning of computers (Dreyfus, 1992, pp. 149–150), Solomonoff's theoretical works in the 1960s had already indicated the vast potentials of machine learning. The significance of his works only started to garner more attention after the recent breakthrough in deep learning, which has led to a proliferation of big data applications.[67] Understanding Solomonoff's proofs will in fact help us grasp the potentials and limits in deep learning.

---

[65] A blind search procedure, such as a depth-first search or breath-first search, explores each choice without considering the type of problem being addressed. A heuristically informed search can improve its search efficiency "if there is a way to order the choices so that the most promising are explored earliest" (Winston, 2000, p. 70). For instance, a greedy search (best-first search) always selects the path that appears most promising at the moment. At each step, it uses a heuristic function to choose the most promising node to expand and explore (Winston, 2000, p. 75).

[66] In his seminal paper "Some Studies in Machine Learning Using the Game of Checkers" (Samuel, 1959), Samuel identifies two general approaches to the problem of machine learning. One is the Neural-Net approach, which can lead to the development of general-purpose machines but "we have a long way to go before we can obtain practical devices," because the size of the artificial neural network, functionable at the time, was much smaller than "the size of neural nets used by animals." Thus he took the second approach, which is a special purpose program that can run more efficiently and is "capable of realization" at his time.

[67] As Solomonoff reflects in the 2005 Midwest NKS (A New Kind of Science) Conference (Adrian German, 2018), his works did not draw much attention when they were first published, and then all of a sudden, there has been a huge amount of interest in his works since the turn of the century.

Therefore, formulating a realistic critique of AI and machine learning would necessitate an understanding of Solomonoff's proofs, which I will present and discuss in the following chapter.

4.5. Summary

This chapter traces through the nuances in articulating the relations between machine learning and cybernetics. The original notion that machines can "learn" can be attributed to the respective research by Wiener, Shannon, and Ashby. Wiener's guided missiles can be perceived as machines capable of learning the flying patterns of aerial targets and predict where they may be at a specific point in time. His concept of negative feedback is the basic scheme for training a machine-learning model today. Shannon's electromechanical mouse can learn from past mistakes in solving a maze such that subsequent trials will take less time. Ashby's homeostat learns from environmental feedbacks to maintain homeostatic stability. All these cybernetic inventions, capable of learning in their own way, are predecessors to machine learning in AI.

Yet, just as AI can be differentiated from cybernetics in its emphasis on simulation and abstraction over scientific knowledge and replications of living mechanisms, the same is true for machine learning. We can already see this differentiation in Turing's two papers on machine intelligence, which proposes the design of "unorganized machines" to explore the possibility of "learning" in a universal Turing machine. The differentiation is also evident in Samuel's checkers-playing program, widely recognized as the first machine-learning program that runs on a universal computer. To develop the checkers-playing program, Samuel was experimenting with heuristics that have nothing to do with actual biological or physiological processes. Solomonoff's proofs on Algorithmic Probability and universal inductive inference are also unrelated to scientific knowledge about mechanisms in a living body. These proofs on the theoretical limitation of using a universal computer to discover data patterns did not garner much attention during the early days of AI, but in recent years, with the advent of big data and deep learning, people have begun to recognize their significance to the research direction of machine learning. In the next chapter, I will attempt to explain Solomonoff's proofs without going into the formal mathematics of the original proofs. I will then show why these proofs are relevant to deep learning and why they anticipate the practice of big data.

Chapter 5.

What Machine Learning Can Learn

Me	Can you explain the math behind the convergence of algorithmic probability?
ChatGPT:	The convergence of algorithmic probability refers to the property that the algorithmic probability of a sequence converges to the true probability of the sequence as the length of the sequence increases. The algorithmic probability of a sequence is defined as the sum of the probabilities of all binary strings that generate the sequence, where the probability of a binary string is determined by its length and the number of strings of that length.

5.1. Overview

In the preceding chapter, I argue that cybernetics is the precursor for machine learning, which departs from cybernetics in its emphasis on abstract computer algorithms over the symmetry of mechanisms between living organisms and machines. Among the pioneering works on machine learning from its early days, Ray Solomonoff's theories on Algorithmic Probability and universal inductive inference have the most significance with regard to the field's recent development. His proofs are formulated in formal mathematical treatises, thus difficult to grasp for anyone not well-versed in probability calculus. Nevertheless, without grasping the central ideas of these proofs, it is difficult to develop a realistic projection of what AI empowered by deep learning is capable of. Acquiring an intuitive grasp of these proofs can help us verify Kai-Fu Lee's claim that AI, with the breakthrough in deep learning, has become better than humans "in its ability to detect incredibly subtle patterns within large quantities of data" (K.-F. Lee & Chen, 2021, p. 430), and that this capability will continue to improve over the coming decades. Lee's claim has been made primarily based on what he has seen in research laboratories and start-up ventures around the world. With his privilege position in the field, he certainly has an authoritative voice on the matter. But from the point of view of those who have not been exposed to the cutting-edge development in AI, it is difficult to gauge whether Lee's opinions on AI's future capabilities are overly optimistic or not. As I will explain in this chapter, grasping the intuition behind Solomonoff's proofs can indeed attest to Lee's claims and opinions.

This verification can help dispel critiques of AI based either on unrealistic projections of its potential or on a lack of appreciation of what it will soon be capable of doing. We ought to differentiate the enthusiasm of AI researchers from the fascination with artificial life that has been deeply ingrained in human culture, from the Talos in Greek mythology[68] to Mary Shelley's *Frankenstein* in the 19th century. AI researchers are not just fabricating stories but trying to engineer concrete realities, constrained by the affordances of existing technologies. They have to face the uncertainty of wasting their research careers away on projects that yield no practical result. Yet, they have remained enthusiastic over their projects on machine intelligence because their vision is supported by computational theories that have been articulated in formal, mathematical proofs. This is true for early AI pioneers, whose conviction in their research largely came from their understanding about the universal Turing machine (UTM) and the Church-Turing thesis. This is also true for researchers in machine learning today, as AI's capability to discover any subtle patterns in data has gained credence within the AI community from formal computational theories articulated in mathematical proofs. These proofs were formulated a long time ago by Ray Solomonoff, who conceptualized the model of Algorithmic Probability and proved mathematically that, given the right circumstances, this model would inductively converge to the true probability behind an actual stochastic process in the real world.

In Section 3.4, I elaborated on the Church-Turing thesis and how Alan Turing's seminal papers on machine intelligence (1948, 1950) inspired a generation of AI pioneers. This chapter draws a parallel between the Church-Turing thesis and Solomonoff's Algorithmic Probability. In much the same way that a UTM can theoretically solve any describable problem, Solomonoff developed an algorithm for a UTM capable of discovering all regularities in a body of data. This chapter surveys the intellectual background of Solomonoff and explains how different philosophies and theories contribute to his conceptualization of Algorithmic Probability. It then extracts the core ideas from Solomonoff's technical papers into everyday language to explicate his model of Algorithmic Probability and how this model anticipates the recent breakthrough in

[68] See footnote 4 in Introduction.

deep learning and big data. Besides offering high-level explanations of Solomonoff's works, this chapter goes into more technically advanced discussion in Section 5.5, which illustrates possible applications of Algorithmic Probability to problems typically addressed in machine learning, such as image classification. Understanding this technical discussion requires basic knowledge of machine learning, and Section 5.2 presents the general framework of how machine learning operates today. As Sections 5.2 and 5.5 presuppose familiarity with algorithmic thinking and some basic knowledge of algebra, readers not well-versed with such a technical mindset may elect to skip these two sections and they should still be able to understand the central ideas behind Algorithmic Probability and its significance from reading other sections.

5.2. A Crash Course on Machine Learning

Gaining some basic knowledge about the way machine learning operates can help us appreciate the significance of Solomonoff's theories. As machine-learning experts Michael Jordan and Tom Mitchell explain, "machine learning algorithms can be viewed as searching through a large space of candidate programs, guided by training experience, to find a program that optimizes the performance metric" (2015). To illustrate what this explanation means, I will go through two applications of machine learning: housing price prediction and image classification.[69]

Let us begin with the application of machine learning to housing price prediction. Suppose we want to come up with a computer program that predicts the market price of a house in Vancouver. One possible way of modeling this real estate market is to hypothetically assume the size of a house as linearly correlated to its market price. Given a historical dataset of sizes and prices of houses sold in Vancouver, we can plot the data as crosses on a graph with the vertical axis denoting the price and horizontal axis denoting the size (see Figure 3). We can then model the relationship between sizes and prices as a straight line through the crosses. There can be many possible such lines, some fitting the data better than others. Each possible line through the crosses is called a "hypothesis," and all such possible lines constitute the "hypothesis space." In general, a machine learning hypothesis is "a candidate model that approximates a target

[69] The examples are adapted from Andrew Ng's course on machine learning. (See https://youtube.com/playlist?list=PLoR5VjrKytrCv-Vxnhp5UyS1UjZsXP0Kj)

function for mapping inputs to outputs" (Brownlee, 2019). This definition is derived from a scientific hypothesis, which can be defined as "a provisional explanation for observations that is falsifiable" (Brownlee, 2019), and from a statistical hypothesis, which is "an explanation about the relationship between data populations that is interpreted probabilistically" (Brownlee, 2019).[70] The so-called learning for a machine learning algorithm "involves navigating the chosen space of hypothesis toward the best or a good enough hypothesis that best approximates the target function" (Brownlee, 2019). It is "a search through the space of possible hypotheses for one that will perform well, even on new examples beyond the training set" (Brownlee, 2019).

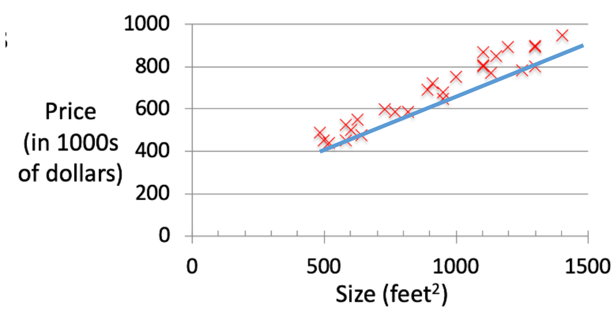


Figure 3: Plotting the data on a "size" vs "price" graph.

Figure 4 shows what happens when a machine learning algorithm tunes its parameters to find an optimal solution. Each line in the three graphs is a particular hypothesis of how we may predict the price based on the size of the house. It can be expressed as a linear function between the price and the size of a house, $Price = \theta_0 + \theta_1 \times Size$. A machine learning algorithm automatically finds the optimal solution by minimizing the gaps between the actual data and the predicted data based on the hypotheses (see Figure 5). Then, through a method called "gradient descent," a machine-learning algorithm would iteratively update the θ_0 and θ_1 parameters, thus gradually moving the hypothesis closer to the optimal hypothesis. For instance, the parameters (θ_0, θ_1) may begin at $(900, -0.4)$, then updated to $(-300, 1.2)$ after one iteration, then to $(100, 0.4)$ after a second iteration. The implication of such tuning of parameters is illustrated in Figure 4. The algorithm starts with a hypothesis of $Price = -0.4 \times Size + 900$ (the graph on the left), then changes to the second hypothesis of $Price = 1.2 \times Size - 300$ (the graph in the middle), before finally moving to the third

[70] The article "What is a Hypothesis in Machine Learning?" (Brownlee, 2019) elaborates on the relations and distinctions between these three definitions of hypothesis.

and optimal hypothesis in $Price = 0.4 \times Size + 100$ (the graph on the right). In general, a machine-learning algorithm iteratively updates its parameters to reduce to total error between the hypothesis and the actual data in a way that approaches an optimal hypothesis.

This simple linear model can be made more complex by factoring in more variables (e.g., the age of the construction, the number of bedrooms, the number of bathrooms, the quality of schools in the neighborhood, etc.), or by making it a non-linear polynomial (e.g., $Price = a + b \times Size + c \times Size^2 + d \times Size^3$). In general, the more complex the model, the easier it is to minimize the distance between actual data and the hypothesis because curves rather than straight lines can be drawn to fit through the data (see Figure 6). But the model may not predict well if it does not capture some actual pattern or rationale behind the training data. To affirm that the model can predict well, a data scientist can test the performance of a hypothesis by computing the total sum of errors for new data examples that do not belong to the training set.

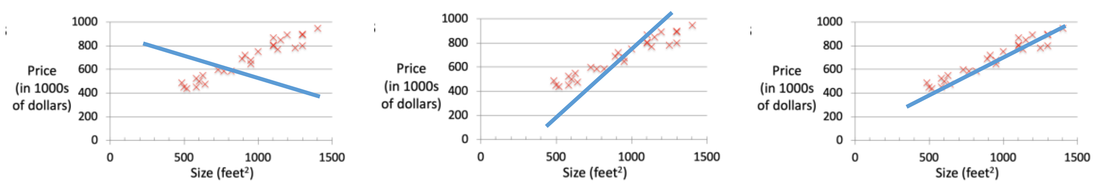
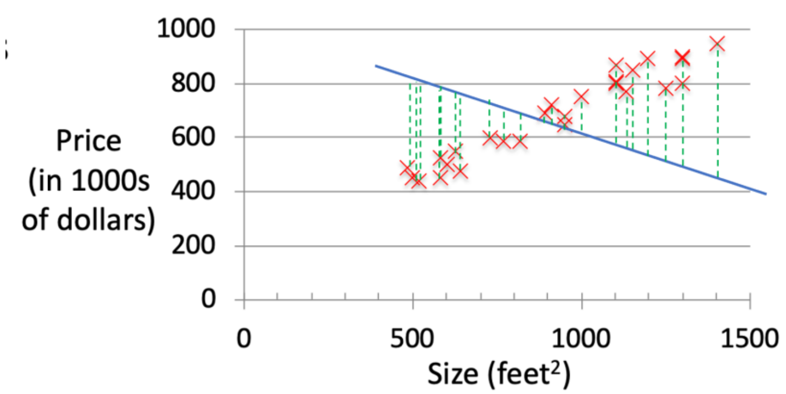


Figure 4: Different hypotheses in a linear regression model



Difference between hypothesis and actual data $= \left(Price_{actual}\left(size_{i^{th} = sample}\right) - Price_{predicted}\left(size_{i^{th} = sample}\right)\right)^2$

Figure 5: Accumulated differences between actual data and predicted data based on hypothesis.

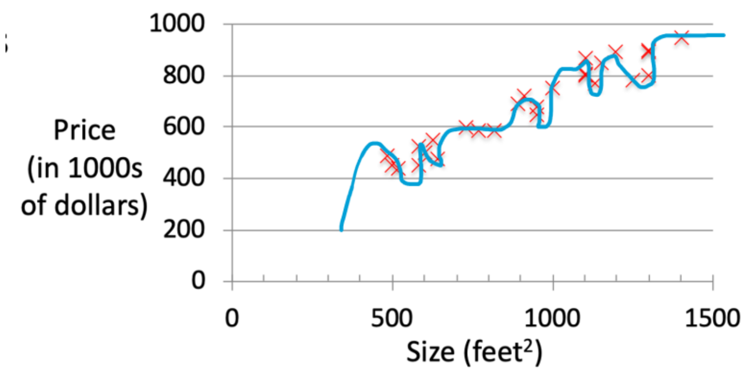


Figure 6: A non-linear regression model that fits the data better but would not predict well.

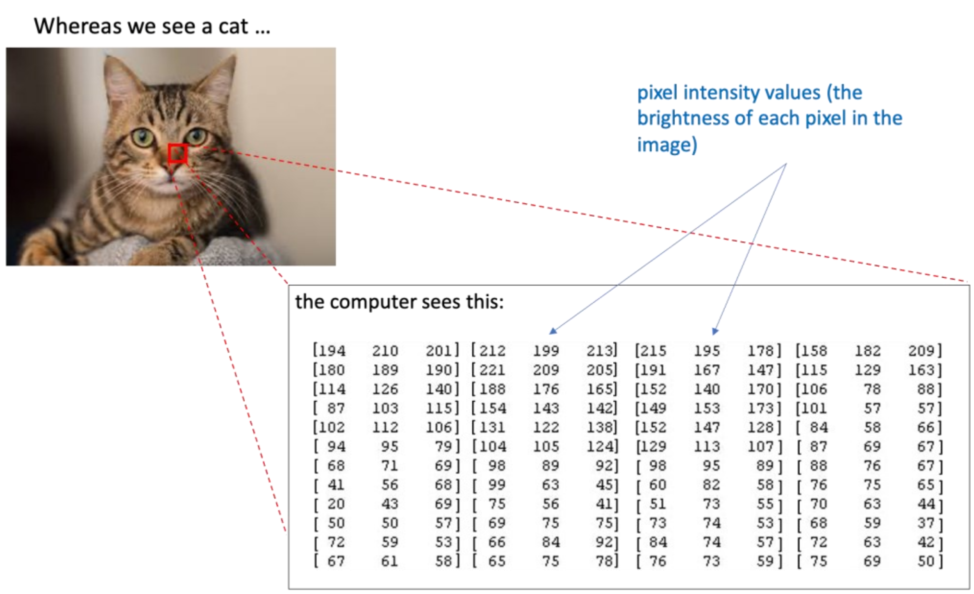


Figure 7: Image as a matrix of pixels[71]

Now, let us look at a more complex example, a machine-learning model that identifies images that contain cats. Suppose we have a training set with half a million images of cats and half a million images with no cat. We want the machine-learning algorithm to find a candidate program that, for most of the images in the training set, correctly classifies whether they contain cats or not. To translate this problem into a machine-learning problem, each image is first converted into a matrix of pixels, which is essentially a grouping of three numbers ranging from 0 to 255 (See Figure 7). A program cannot "see" an image as it appears to human eyes. It can only process a digital representation of an image as a vector of numbers. What we want is a program that

[71] The image in this figure is taken from the web (Shirodkar, n.d.).

takes a training set of numeric vectors as inputs and yields correct results of 1's (for images with cats) and 0's (for images with no cat) (See Figure 8). Suppose "Candidate Program A" and "Candidate Program B" are tested against a collection of data examples different from those used to train the programs, and "Candidate Program A" is correct for 99% of the cases and "Candidate Program B" is correct for only 65% of the cases. Then we would deem "Candidate Program A" as the better hypothesis. A machine learning algorithm would presumably iterate through different candidate programs from the hypothesis space of a chosen model, such as an artificial neural network or a support vector machine and pick out the candidate program that classifies more correctly than others. One hypothesis that can be correct 100% of the time is a program with a million if-statements (see Figure 9), but this program would be unable to test whether a new image contains a cat because it has not uncovered any underlying regularities or patterns about the pixels that represent a cat image. Hence, an ideal candidate program is both accurate on classifying the training data and has a relatively simple logic, because a simpler logic implies a better generalization about the underlying pattern of pixels that corresponds to cat images.

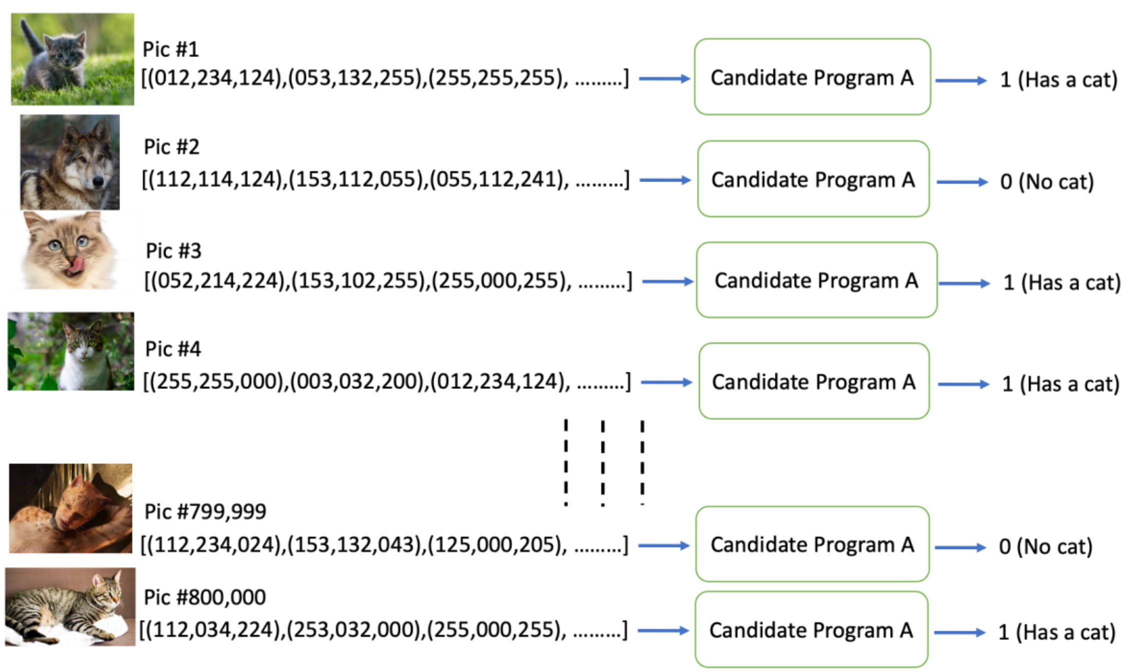


Figure 8: Training a machine-learning model to recognize pictures of cats.[72]

[72] The images in this figure are taken from the web (Allen, n.d.; Forest, 2016; *Influenza in Cats | CDC*, 2024; Roberts, n.d.; Shirodkar, 2023; Zielinski, 2020).

The linear regression model is not suitable for modelling a classification problem. Instead, a data scientist would likely choose between logistic regression, support vector machine or artificial neural network as the machine-learning model to be trained. Each model represents a certain class of candidate programs. Between these three models, the time to train (i.e., the time to find a candidate program with good performance) and the accuracy of the selected candidate programs (i.e., the accuracy of the trained machine-learning models) in recognizing cat images would be different. For instance, for simple classification problems, it takes more time to train a neural network than a logistic regression model. But for image classification, in which the number of input features is very large because each pixel is mapped to one feature, training an artificial neural network would be more efficient than training a logistic regression model.

```
If input == [(012,234,124),(053,132,255),(255,255,255), .........]
then return 1
else If input == [(112,114,124),(153,112,055),(055,112,241), .........]
then return 0
else if input == [(052,214,224),(153,102,255),(255,000,255), .........]
then return 1
else if input == [(255,255,000),(003,032,200),(012,234,124), .........]
then return 1
                     |    |    |
                     |    |    |
                     |    |    |
else if input == [(112,234,024),(153,132,043),(125,000,205), .........]
then return 0
else if input == [(112,034,224),(253,032,000),(255,000,255), .........]
then return 1
```

Figure 9: Hardcoding a Correct Program

Having some basic knowledge about artificial neural network would be helpful for understanding the rest of this chapter. A neural network is a key component in the transformer model designed for natural language processing such as language translation.[73] Such problems, as Dreyfus pointed out, are beyond the affordance of symbolic AI based on formalizable rules (see Sections 1.3 and 1.4). The design of an artificial neural network is originally motivated by the goal of building machines that mimic the biological model of a brain. In this model, a brain consists of many

[73] An overview of the transformer model is presented in the Appendix.

interconnected neurons. Each neuron is a brain cell, which has a few input wires called dentrites and a single output wire called axon. If the sum of the signals accumulated from the dentrites crosses certain threshold, the neuron would fire a signal to other neurons through its axon. For instance, when a needle pricks your skin, the nerves on your skin would transmit signals to the dentrites of specific neurons in your brain, triggering them to fire signals along their axons that would ultimately lead to bodily movements. Modelling this behavior, an artificial neuron is a computational unit that takes in several inputs through input wires, performs some computations, and sends the computed result through its output wire to other artificial neurons (see Figure 10). The computation performed by an artificial neuron is typically the sigmoid function, which returns a number close to 1 if the weighted sum of the input values is above certain threshold.[74] An artificial neural network is simply a network of interconnected artificial neurons, which are usually called “network nodes.” It is possible to connect network nodes in many ways, but in a typical neural network architecture, the nodes are arranged in layers. For instance, Figure 11 shows a network with four layers. The first layer, the input layer, has 3 nodes. The input to the model would be a vector of 3 numbers, with each number passed to one of the nodes on the input layer. Each of Layer 2 and Layer 3 has 5 nodes. The output layer has four nodes, meaning that the output of this neural network is an output vector with four numbers. In general, a neural network with more layers and nodes is more complex and can achieve better optimization metrics. The trade-off is that a complex neural network needs to be trained with massive amount of data to predict well for real-life data, and such a training operation takes a long time to complete.

[74] Referring to the sigmoid function in Figure 10, the effect of the function can be explained as follows: (1) If the variable (z in the above diagram) is a large enough number, the function would return a number close to 1. In the above diagram, if z is greater than 4, the sigmoid function would return 0.99 or greater. (2) If the variable (z in the above diagram) is a small enough number, the function would return a number close to 0. In the above diagram, if z is less than -4, the sigmoid function would return 0.01 or smaller.

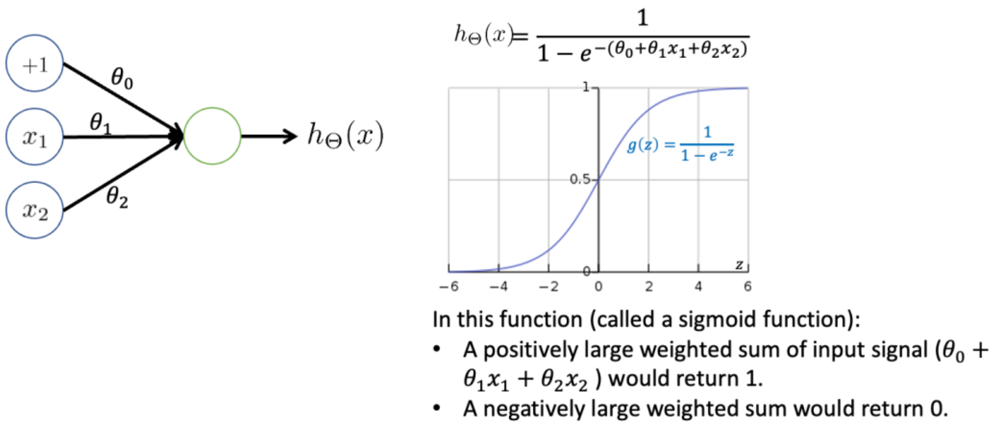


Figure 10: An artificial neuron

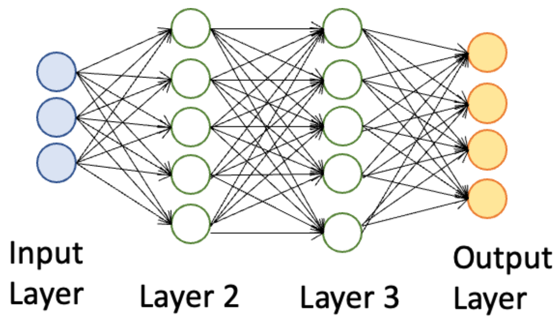


Figure 11: A typical neural network architecture

The examples of housing price prediction and cat image classification illustrate how a real-life problem can be represented as numerical inputs to a machine-learning model that can be automatically tuned to generate some expected outputs (see Figure 12). The training data would be a collection of such inputs and expected outputs. Once data scientists have come up with such an abstract representation of a real-life problem and collected the training data based on this representation, training a machine-learning model is simply a matter of “searching through a large space of candidate programs" and find one that “optimizes the performance metrics” (Jordan & Mitchell, 2015). Performance metrics is typically optimized by minimizing the differences between the outputs generated by the machine-learning model and the expected outputs based on the training data.

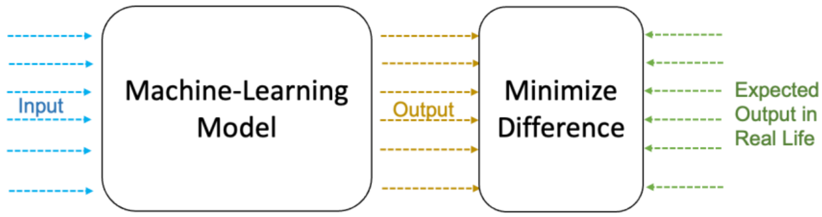


Figure 12 A Generic Framework for Machine-Learning Model

But what gave the computer scientists and AI researchers the confidence that massive amount of data combined with ever more powerful CPU's (central processing units) would make it possible to train machine-learning models that optimize performance metrics? Researchers in computational theories play a key role of substantiating speculative hunches with formal theoretical proofs. As I will explain next, prior to the recent empirical successes in deep learning, the concepts of big data and deep learning have already been mathematically proven by Ray Solomonoff in the early days of AI research.

5.3. The Intellectual Background for Solomonoff's theories in Machine Learning

In the introduction to Ray Solomonoff 85th Memorial Conference, David L. Dowe hails Solomonoff as "the original pioneer, in the early 1960s, of the use of (universal) Turing machines (using algorithmic information theory and Algorithmic Probability) for prediction problems in statistics, machine learning, econometrics and data mining" (Dowe, 2013, p. 2). In this section, I will describe what led to Solomonoff's discovery of Algorithmic Probability. I first explain the influence of Rudolf Carnap on Solomonoff, particularly the philosopher's multiple definitions of probability and his view that the entire universe can be represented digitally. I then bring up the historical anecdote between John McCarthy and Solomonoff at the Dartmouth summer workshop, in which Solomonoff discovered that inverting a Turing machine is the key to address the deficiency in Carnap's system of probability.

Solomonoff describes his "voyage" toward discovering this algorithmic method in the paper "The Discovery of Algorithmic Probability" (1997). Algorithmic Probability is a method that provides a measure of the inherent randomness of a sequence of symbols. The method defines the "prior probability" assigned to the sequence and uses Bayes' Theorem to find the probability of a particular continuation of the sequence. "Prior probability," also interchangeably called "a-priori probability," is the technical term for the unconditional probability used in Bayes' theorem. It represents the probability of a hypothesis before any condition, such as the collection of data, is taken into account. As an undergraduate at the University of Chicago, Solomonoff was studying philosophy from Carnap, who at the time "was working on a general theory of probability as much as [he] was" (R. J. Solomonoff, 1997, p. 6). Solomonoff got two important ideas from Carnap: (1) multiple definitions of the term probability, and (2) a "long sequence of symbols"[75] as "a description of the entire universe" (R. J. Solomonoff, 1997, p. 6). From Carnap, "[t]he best known was the frequency concept of probability," but there is another definition of probability, which is "the degree of confidence one had in a hypothesis with respect to a certain body of data" (R. J. Solomonoff, 1997, p. 6). For instance, if you are to pick a ball out of a bag of 500 blue balls and 9500 red balls, the prior probability of picking out a blue ball is 5 percent. But suppose you have no prior knowledge about the bag of balls. If you have picked out a sample of 100 balls, 8 being blue and 92 being red, the frequency of a blue ball in this sample is 8%. We may estimate that the next ball you pick has a probability of 8% being a blue ball. This frequency interpretation of probability is subjective because it is based on observations conducted subjectively. It is an approximation of the prior probability, but there is a significant gap between 8% and 5%. Nevertheless, the more balls you pick out, the closer is the probability estimation to the prior probability, and correspondingly, the greater is the degree of confidence or confirmation in your hypothesis.

As Anna Longo explained in *Le jeu de l'induction* (2022), Carnap wanted to evaluate the logical implication between the results of a repeated experiment (the frequencies observed) and the probability of future observations. This measure depends on the strength of the implication between the premises and the consequences of a probabilistic inference, in particular the logical relationship between the proposition that describes the evidence collected and the predictive hypothesis (Longo, 2022, pp. 24–25).[76] For instance, if a dice is made up of malleable material like paper, the shape of

[75] As I will explain in Section 5.4, a symbol in this long sequence, for Carnap and for Solomonoff, can be anything from an integer (in a number sequence) to an English word (in a sentence) to an image of a cat (in a corpus of cat images).

[76] "Rudolf Carnap propose d'évaluer l'implication logique entre une proposition portant sur les résultats d'une expérience répétée (les fréquences observées) et celle qui exprime la probabilité des observations futures. L'induction devient ainsi une question de quantification du degré de confirmation des hypothèses sur la base de leur implication logique. Il est important de souligner

this paper dice would become malformed after repeated experiments. Unlike throwing a well-formed dice, the frequency observed in repeatedly throwing this dice would not approach the probability of $1/_6$ as the predictive hypothesis. This illustrates why the probability of future observations cannot always be inferred from the frequency observed. Carnap thus aims to determine the degree of confirmation of a hypothesis by considering it as an inference whose validity can be established based on the axioms of inductive logic. Consequently, scientific theories are propositions that can ideally be deduced analytically from a number of atomic statements that relate to fundamental empirical data.[77] Carnap's inductive logic presumably makes it possible to calculate the degree of objective belief that can be attributed to any inductive hypothesis (Longo, 2022, p. 25).[78] As Longo explains, Carnap's attempt to analytically deduce true propositions from basic axioms is problematic because it would be impossible to explain how we manage to establish these first definitions without falling into an infinite regression (Longo, 2022, p. 26).[79] After conceding the impossibility of fixing the axioms

---

la distinction fondamentale de Carnap entre l'estimation de la probabilité d'un événement selon une hypothèse, définie ⟨⟨probabilité 2⟩⟩, et la probabilité de l'hypothèse prédictive, ou ⟨⟨probabilité 1⟩⟩, cette dernière mesure son degré de confirmation. La *probabilité* 2 se calcule à partir des fréquences observées par répétition d'une expérience scientifique, et est une prédiction de la probabilité des observations futures. En revanche, la mesure de la *probabilité* 1 dépend de la force de l'implication entre les prémisses et les conséquences d'une inférence probabiliste, notamment la relation logique entre la proposition qui décrit les éléments de preuve collectés et l'hypothèse prédictive" (Longo, 2022, pp. 24–25).

[77] In Carnap's framework, atomic statements are basic statements or propositions that cannot be further decomposable into simpler statements. They represent the fundamental building blocks of knowledge and are often used as the starting point for logical reasoning. In the context of inductive logic, atomic statements are treated as basic statements or propositions about which probabilities can be assigned. Carnap's axioms allow us to reason about the probabilities of atomic statements and make inferences based on available evidence. Using the axioms, we can combine atomic statements and their probabilities to derive logical conclusions about more complex statements, such as compound statements or hypotheses (see *Protocol Sentence | Logic, Meaning & Truth | Britannica*, n.d.; *Rudolf Carnap > C. Inductive Logic (Stanford Encyclopedia of Philosophy)*, n.d.).

[78] "Carnap vise ainsi à déterminer le degré de confirmation d'une hypothèse en la considérant comme une inférence dont la validité peut être établie en s'appuyant sur les axiomes de la logique inductive. En conséquence, les théories scientifiques sont des propositions qu'on peut idéalement déduire de façon analytique à partir d'un certain nombre d'énoncés atomiques qui portent sur les données empiriques fondamentales. Lorsqu'une théorie confirmée est une inférence nécessairement vraie, en revanche, une hypothèse probable est admissible, mais partiellement impliquée par les observations disponibles. La logique inductive de Carnap était ainsi censée permettre de calculer de façon précise le degré de croyance objective à attribuer à toute hypothèse inductive" (Longo, 2022, p. 25).

[79] "D'après Quine, il serait cependant impossible d'expliquer comment l'on arrive à établir ces définitions premières sans tomber dans une régression infinie, et les paradoxes qu'elle entraîne" (Longo, 2022, p. 26).

of a universal language of empirical sciences, Carnap admits a plurality of conventional logics relating to the various fields of scientific enterprise (Longo, 2022, p. 25).[80]

As explained in “The Discovery of Algorithmic Probability” (1997), Solomonoff wanted to overcome the difficulties associated with Carnap’s plurality of conventional logics, and this theoretical endeavor eventually led to his invention of Algorithmic Probability. On the one hand, he adopted Carnap’s representation of the universe as a long sequence of symbols. He was also influenced by how Carnap “was able to assign a priori probabilities to any possible string of symbols that might represent the universe” and derived the degree of confirmation from these a priori probabilities using Bayes’ Theorem (R. J. Solomonoff, 1997, pp. 6–7).[81] On the other hand, he found Carnap’s method of computing prior probabilities unreasonable because the “probability distribution depended very much on just what language was used to describe the universe,” rather than directly on the data (1997, p. 7).

Solomonoff’s was working on overcoming the weaknesses in Carnap’s method when he attended the Dartmouth summer workshop. It was from a casual conversation with John McCarthy that Solomonoff discovers Algorithmic Probability, which reframes Carnap’s probabilistic universe into a theory based on UTM. During one of the sessions in the Dartmouth summer workshop, McCarthy was giving a talk on the thesis that “all mathematical problems could be formulated as problems of inverting Turing machines” (R. J. Solomonoff, 1997, p. 8).[82] It occurred to Solomonoff that McCarthy’s thesis might

---

[80] “Dans la première édition de *The Logical Foundations of Probability*, Carnap était persuadé qu’il était possible de calculer le degré de confirmation de théories différentes par une fonction unique à partir des axiomes de la logique inductive. Cependant, une fois établie l’impossibilité de fixer les axiomes d’un langage universel des sciences empiriques, il admet alors une pluralité de logiques conventionnelles relatives aux divers champs de l’entreprise scientifique” (Longo, 2022, p. 25).

[81] Solomonoff’s abstract description of Carnap’s model becomes clearer with an understanding of Algorithmic Probability (see Section 5.4). For instance, consider a corpus of digital images organized in a long sequence. Carnap’s model would assign to each image in the long sequence an a-priori probability, which may indicate the chance that there is a figure of a cat in the image. Using Bayes’ Theorem, these a-priori probabilities can be used to compute the conditional probability that there is also a figure of a cat in a new image, given the observations already made on the images from the corpus. The computed conditional probability is in fact the degree of confirmation.

[82] We can define a Turing Machine as a function M that takes in an input string p and produces an output string s. In other words, we are given the machine M and an input string p, and the two combines to give an output string s. Formally, this can be represented as $M(p) = s$. Conversely, inverting a Turing machine can be defined as follows: We are given a machine M and a desired

solve his induction problem. He asked McCarthy, “Suppose you are given a long sequence of symbols describing events in the real world. How can you extrapolate that sequence” (R. J. Solomonoff, 1997, p. 8)? McCarthy’s response confirmed Solomonoff’s intuition: “Suppose we were wandering about in an old house, and we suddenly opened a door to a room and in that room was a computer that was printing out your sequence. Eventually it came to the end of the sequence and was about to print the next symbol. Wouldn’t you bet that it would be correct” (R. J. Solomonoff, 1997, p. 8)? In other words, if there exists a program on a UTM that can accurately predict your sequence, it must have identified the logical function for generating the sequence.[83] This logical function can then be used to predict what comes after the sequence. Finding this logical function can be treated as a problem of inverting a Turing machine in which the logical function is represented as an input string to the Turing machine. This conversation with McCarthy marks the moment of eureka for Solomonoff and led to his discovery of Algorithmic Probability, an algorithmic method that provides “a useful estimate of a sequence’s true probability of being outputted by” a Turing machine (Campbell, 2013). This application of UTM allows Solomonoff to turn Carnap’s philosophical model of prior probability distribution into an actual computing algorithm without stipulating language-dependent axioms. In the next section, I will bypass the mathematical details in the formal proof by Solomonoff and explain how the pseudo algorithm of Algorithmic Probability can identify all the subtle patterns in massive amount of data.

5.4. An High-Level Explanation of Solomonoff’s Algorithmic Probability

Solomonoff’s seminal papers on Algorithmic Probability and universal inductive inference are primarily mathematical and would elude the minds of those who are not well-versed in the mathematical reasoning of conditional probability. Nevertheless, if his reasoning can be described in non-mathematical everyday language, those outside the discipline of computational theory of machine learning can gain a better sense about the

---

output string s, and the problem is to find a string p such that $M(p) = s$. (R. J. Solomonoff, 1997, p. 8).

[83] It is of course possible that the logical function is simply the printing out of the exact same sequence, in which case the prediction would likely be false. This is where the introduction of Occam Razor comes in, as the shortest and simplest program would be preferred over long and complicated programs.

significance of his works. The understanding resulted from this endeavour is especially crucial today with the advent of deep learning and big data, whose vast potentials and scope of capability can be anticipated by Solomonoff's proofs.

Solomonoff invented Algorithmic Probability by combining Carnap's ideas and inverting Turing machines with Epicurus's principle of multiple explanations and Occam Razor (Li & Vitányi, 2008, p. 347). He first describes the algorithm in "A Preliminary Report on a General Theory of Inductive Inference" (1960). As he explains, the goal of the algorithm is "concerned primarily with the problem of extrapolation of a very general time series, whose members may be numbers or non-numerical objects, or mixtures of these" (1960, p. 1). Following Carnap, this "very general time series" can be represented as "a very long sequence of symbols," which may be "a passage of English text, or a long mathematical derivation" (R. J. Solomonoff, 1960, p. 1). In other words, the goal is to uncover any underlying regularities or patterns behind a time series of symbols. If each symbol is an English word, the underlying regularities would be the English grammar (R. J. Solomonoff, 1997, p. 12). If each symbol is some digital image of a cat, the underlying regularities would be the distinction of pixels unique to images of cats. Solomonoff models this problem of extraction with a universal Turing machine that takes "any finite string of 0's and 1's as acceptable input" and produces a sequence of symbols as output. Solomonoff defines such an input bitstring as a description of the output sequence of symbols with respect to the UTM, and "[i]n general, any regularity in a corpus may be utilized to write a shorter description of that corpus" (R. J. Solomonoff, 1964, p. 8).[84]

Whereas Carnap attempts to derive a "true" and "absolute" prior probability of a sequence, Solomonoff avoids the same pitfall by inventing an algorithm that is only concerned with relative probabilities. Algorithmic Probability is a particular framework that defines how each sequence is assigned a certain a-priori probability. Each assigned probability can be viewed as a measure on the inherent randomness of the sequence. A sequence with a higher assigned probability is more likely to contain patterns and regularities in the sequence. The exact value of the probability assigned to each sequence is not very meaningful. But the relative difference in probabilities is sufficient

[84] Section 5.5 explains why "any regularity in a corpus may be utilized to write a shorter description of that corpus."

for Solomonoff to calculate the probabilities of particular continuations of the sequence, where the continuation is generated based on patterns or regularities identified in the sequence. This framework allows Solomonoff to use Bayes' Theorem to prove that the continuation of a sequence would converge with the stochastic function of an underlying phenomenon if more and more data are collected to extend the length of the sequence. As Longo explains, by relying on relative probabilities in his proof, Solomonoff implicitly assumes Bayesian subjectivism, which allows him to overcome the main issue with Carnap's project: the impossibility of establishing an ideal language to measure the probability of hypotheses in a perfectly objective way (2022, pp. 48–49).[85]

Formally, Solomonoff defines Algorithmic Probability as the sum of the probabilities for all inputs that can produce the target output on the UTM. The probability of generating every specific program is 2^{-n_i} where n_i is the number of bits in i^{th} program, and Algorithmic Probability is equal to $\sum_{i=1}^{\infty} 2^{-n_i}$. Since the program with the shortest length will have a probability several magnitudes higher than the other programs (e.g., a program with eight fewer bits is already $2^8 = 256$ times less likely), the Algorithmic Probability is approximately the probability of the shortest program. Because a shorter program is likely simpler than a longer program, Solomonoff's algorithm also satisfies Occam's Razor, which states that "all things equal, explanations that posit fewer entities, or fewer kinds of entities, are to be preferred to explanations that posit more."[86]

The mathematical proof that Algorithmic Probability would converge to the probability of some underlying stochastic function came a number of years after the invention of the algorithm. At a keynote given at the Midwest NKS (A New Kind of Science) Conference in 2005 (Adrian German, 2018), Solomonoff shared his reflections on his early works. When he first introduced Algorithmic Probability in "A Preliminary Report on a General Theory of Inductive Inference" (1960) and in "A Formal Theory of Inductive Inference" (1964), he felt satisfied with the intuition behind Algorithmic

[85] "C'est le subjectivisme bayésien qui permet à Solomonoff – contrairement à son professeur Rudolf Carnap – de ne pas se soucier de l'impossibilité d'établir un langage idéal permettant de mesurer la probabilité des hypothèses d'une manière parfaitement objective. Il considère ainsi le choix du langage de programmation comme impliquant une évaluation subjective de la probabilité des hypothèses à tester, tout en sachant que les estimations devraient converger vers des valeurs similaires grâce au processus de conditionnalisation. La formule de Bayes permet de sélectionner la prédiction à laquelle revient le degré supérieur de confiance, et ce indépendamment du langage de programmation" (Longo, 2022, p. 49).

[86] See www.britannica.com/topic/Occams-razor.

Probability, but was not completely certain that it is correct. A few years later, around 1968, he discovered a mathematical measure to confirm this intuition. This mathematical proof was not published until almost a decade later, in the papers "Inductive Inference Theory — A Unified Approach to Problems in Pattern Recognition and Artificial Intelligence" (1975) and "Complexity-Based Induction Systems: Comparison and Convergence Theorems" (1978).

In the proof, Solomonoff first assumes that the sequence has been generated by some stochastic probability function. He then asks, what is the conditional probability that the next bit is 0 or 1, given the sequence we have now? Algorithmic Probability, as the probability that his algorithm would assign to the next output being 1, would be different from the true conditional probability, as computable by some underlying stochastic function, that the next output bit is 1. Using Bayes' Theorem, Solomonoff develops his proof that the conditional probability computed by Algorithmic Probability converges with the conditional probability of the next bit output by the stochastic probability function. What Solomonoff manages to prove is the following. Suppose the output sequence has m symbols. We can calculate the deviation between the conditional probability of the second bit given the first bit, computed by Algorithmic Probability, and the true conditional probability of the same condition. We can then do the same for the conditional probabilities of the third bit given the second bit, and the fourth bit given the third bit, and so on. Solomonoff proves that the expected value of the sum square of all such deviations is always less than some constant, even as m increases toward infinity[87]. In other words, if we can keep taking more observations (thus increasing m),

---

[87] The mathematical proof itself is a manipulation of mathematical logic that cannot be explained in layman terms. But it may be helpful to examine what he means by the expected value and the formula of his proofs. As Solomonoff describe, the expected value is based on "a model of induction … in which all possible induction models are formally considered. The predictions of each possible induction model are used in a weighted sum to obtain predictions that are at least as "good" (in a certain stated sense) as any of the component induction models" (1964, p. 17). Formally, the formula of the expected value is:

$$E\left(\sum_{i=1}^{m}(\delta'_i - \delta_i)^2\right) = \sum_{k=1}^{2^m}\left(P\left({}^{k}A^{(m)}\right)\right)\sum_{i=1}^{m}\left({}^{k}\delta'_i - {}^{k}\delta_i\right)^2 \leq b\ln\sqrt{2}$$

eventually the expected value will plateau toward a constant. Therefore, the deviation between Algorithmic Probability and the true probability becomes increasingly negligible. This proves mathematically that Algorithmic Probability converges very rapidly to the true probability as the number of observations increases (R. J. Solomonoff, 1975, p. 5).

The most significant property of Solomonoff's Algorithmic Probability is its completeness. As Solomonoff concludes from his theoretical modelling and mathematical proofs, Algorithmic Probability

> is the only induction system we know of that is 'complete.' By this we mean that if there is any describable regularity in a body of data, Algorithmic Probability is guaranteed to discover it using a relatively small sample of the data. It is the only probability evaluation method known to be complete. (1997, p. 21).

This algorithm takes every possible hypothesis into consideration, including all scientific laws and hypotheses scientists have made to date. Hence he claims, his model

> is at least as good as any other model of the universe in accounting for the sequence in question. Other models may devise mechanistic explanations of the sequence in terms of the known laws of science, or they may devise empirical mechanisms that optimumly [*sic*] approximate the behavior and observations of the man within certain limits. Most of the models that we use to explain the universe around us are based upon laws and informal stochastic relations that are the result of induction using much data that we or others have observed. The induction methods used in the present paper are meant to bypass the explicit formulation of scientific laws, and use the data of the past directly to make inductive inferences about specific future events. (R. J. Solomonoff, 1964, p. 16)

> Here E is the expected value with respect to P. ${}^{k}A^{(m)}$ is the kth sequence of length m. There are just 2^m of them. ${}^{k}\delta'_i$ and ${}^{k}\delta_i$ are the conditional probabilities for the ith bit of ${}^{k}A^{(m)}$ for P^M and P, respectively.
>
> The expected value of the mean square error between the conditional probabilities is less than $\frac{b}{m}\ln\sqrt{2}$. (R. J. Solomonoff, 1975, p. 5)

The means, we would consider every possible output sequence of length m (2^m possibilities in total). These output sequences cover all the possible output sequences as we move inductively from $m = 1, m = 2$, and so on to compute the conditional probabilities for each output sequence. For every possible output sequence, we compute the square of the deviation between the two conditional probabilities, one computed by Solomonoff's model, the other being the true conditional probability). We then account for all possible output sequences by summing up all these square differences, weighted by the probability (the true one assigned by the stochastic function) for each possible output sequence.

Therefore, in theory, Algorithmic Probability can replace the scientific method for discovering physical laws from observations. For instance, if the expected output sequence corresponds to observations on the speed and acceleration of free-falling objects on earth, the algorithm would yield some program that takes into consideration the law of gravity along with other variable factors such as friction or air resistance. The program would in fact make better predictions than calculations based on formal scientific laws, which would necessarily oversimplify the body of data and fail to detect subtle regularities.

A second most important property of Algorithmic Probability is its incomputability. It is not computable because of the infamous halting problem associated with universal machines. Some of the randomly generated inputs may turn out to be programs (e.g., an infinite "while" loop) that never halt to yield an output. As Solomonoff explains, "any computable probability measure cannot be complete" because "there have to be regularities that are invisible to" any probability evaluation method (1997, p. 21).To address this halting problem, Solomonoff suggests that in practice we can approximate the Algorithmic Probability by setting a time limit to the running of every program on the universal machine. Relaxing the time limit would increase the precision of the approximation. Even though there is always the possibility of rejecting an ideal program that requires more time to yield an output, this approximation should be useful in practice, just like the mathematical approximations in computing $\sqrt{2}$ or π (2006, p. 6).

While the relaxation of the time limit for running each program would expand the coverage of possible hypotheses, collecting more data for the output sequence would help narrow down possible hypotheses by invalidating those incompatible with new observations. Eventually, if the time limit is set large enough for programs to run, and if enough observations have been made, the algorithm should have considered most of the possible hypotheses, which are also narrowed down to the ones that best fit the data.

5.5. More technical details on Algorithmic Probability

The previous section attempts to explain Algorithmic Probability in everyday language that skip over technical details. Careful readers may wonder, why would descriptions utilizing regularities be necessarily shorter than descriptions without taking regularities into account? In addition, how can algorithmic probability be useful to address a typical machine learning problem, such as image classification? This section attempts to address these questions, but doing so necessarily involve technical sophistication in my arguments. Readers without the necessary technical background may elect to skip this section, as these are peripheral questions to the understanding of algorithmic probability.

To see why descriptions utilizing regularities would be shorter, consider the example as shown in Figure 13 and Figure 14. The output sequence consists of three symbols: “ABCABCABCABCABC,” “GRYGRYGRYGRY,” and “KKWKKWKKW.” In general, we may assume that the input description bitstring follows a format where it begins with a software program and is subsequently followed by a series of input strings intended for that program. As shown in Figure 13, if no regularity is uncovered, the program would simply reprint the entire string it receives as its input. The length of the overall input description would be the length of the program plus the length of “ABCABCABCABCABC,” “GRYGRYGRYGRY,” and “KKWKKWKKW,” which is equaled $3 \times 12 = 36$. If the regularity of reprinting three letters four times is considered, the length of the overall input description would be the length of the program plus the length of “ABC,” “GRY,” and “KKW,” which is equaled to $3 \times 3 = 9$. For a very long output sequence, the length of the program would be negligible and the input description that takes regularity into consideration would be four times shorter.

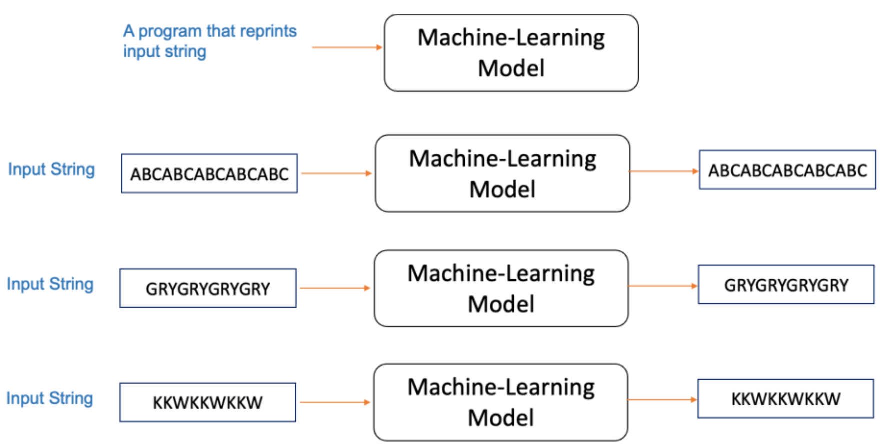


Figure 13: A plain input description of output sequence

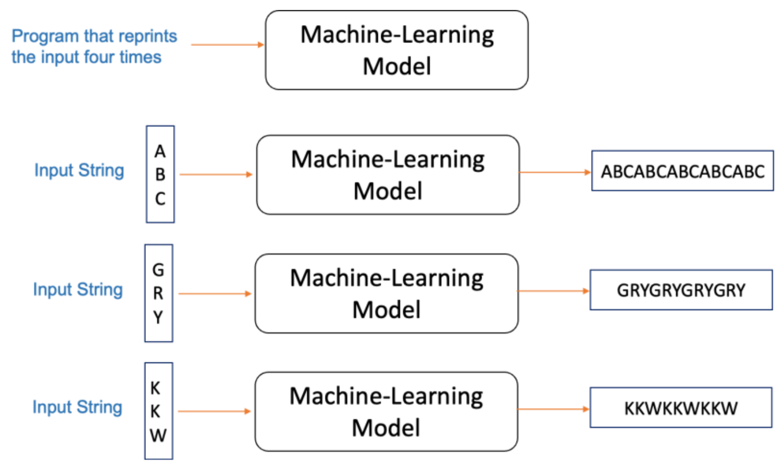


Figure 14: A generalized input description of output sequence

This example illustrates why "any regularity in a corpus may be utilized to write a shorter description of that corpus." It can also illustrate why a very long output sequence can rule out many invalid hypotheses about its underlying regularities. For instance, if the next symbol of the target output sequence is "BCADFEGBCADFEG," then the regularity identified by the program in Figure 14 would no longer be valid. It is evident that extending the output sequence with more symbols can rule out some of the generalized descriptions of shorter output sequences. Therefore, a very long output sequence of symbols can rule out many invalid assumptions about the sequence's underlying regularities.

We can also use the above algorithm to identify regularities in a sequence of complex objects such as digital images, which are essentially vectors of pixels. Consider the problem of recognizing cat images described in Section 5.2. Suppose we have collected a corpus of cat images. The expected output of the UTM would be a long sequence of cat images from the corpus. If no regularity can be identified, the input description could simply be a sequence of digital representations (vectors of pixels) that correspond to the cat images in the output sequence, just like the hardcoding solution in Figure 9. In this case, the length of the input description would be about the same as that of the output sequence. But if some underlying regularities can be uncovered, for instance the identification of a round face, two triangular shaped ears, and whiskers, the description of each image can become simpler by combining this general description with a shorter description specific to each image in the output sequence. The overall input description of the output cat images would be shorter if it is generalized with patterns shared by the images. To pursue this line of reasoning further, if there are more than one underlying patterns in the output sequence, the program that identifies all the patterns would yield the shortest description in comparison to programs that identify a subset of such patterns. Thus, the goal of finding a minimal, shortest description of an output sequence of symbols is equivalent to the goal of discovering all the underlying regularities in the output sequence.

Now, each input description, as a concatenation of a software program and a sequence of input strings, can be represented digitally as a long binary string of 0's and 1's. So, given a randomizer that generates a random sequence of 0's and 1's, it is possible that this randomizer may generate the digital representation of a particular input description. The probability that this randomizer would generate a particular input description would be $(1/2)^{L_{desc}}$, where L_{desc} is the length of the input description. In Algorithmic Probability, a randomizer would iteratively generate 0 or 1 as input to the UTM. If at some point, the machine halts and produces an output that does not match the expected output sequence of symbols, the UTM would be reset. The next bit generated by the randomizer would be treated as the first bit of a new input sequence. After many iterations, many binary input strings would be generated. Most of these input strings will not yield meaningful outputs on the UTM, but some may turn out to be runnable software programs. And because of the universality of a UTM, equipped with the property that any describable procedure can be written as a program on a universal machine, it is possible to express every possible function that describes the regularity of a sequence of symbols as an input to the UTM. If we assume no time constraint and run the algorithm long enough, there is a good chance that the randomizer would eventually generate a bitstring that matches the input description with minimal length. In this method, Solomonoff has adopted Epicurus principle of multiple explanations (Li & Vitányi, 2008, p. 347), which states that all possible hypotheses ought to be considered before narrowing them down to the simplest and most likely explanation.

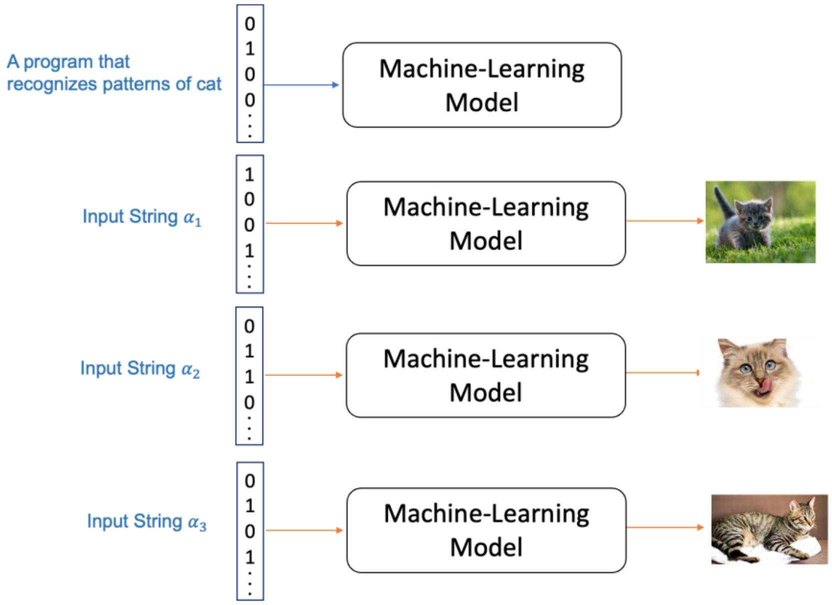


Figure 15: A minimal description of cat image as a concatenation of a program with input strings[88]

This algorithm can be applied to solve the problem of automatic image classification. For instance, given a digital image, can a machine tell whether this image contains a cat or not? In Algorithmic Probability, this becomes a question of whether a new image shares the same regularities that are identified by the minimal-length description about an output sequence of images. As mentioned earlier, the format of this minimal-length input bitstring can be a software program followed by a sequence of inputs that correspond to the output sequence of images. We want to compute the conditional probability for a UTM to generate the new image, if the machine has already processed an input bitstring that corresponds to a minimal description of the output sequence. This is therefore a question of inductive inference. For example, suppose we have collected a corpus of cat images. We are then given a new image and asked to test whether or not this image contains a cat. We can run the algorithm in Algorithmic Probability for a few days and discover some minimal-length input description about a very long sequence of cat images drawn from the collected corpus. We can think of this minimal-length input description as a program concatenated with a sequence of specific input strings (see Figure 15). We can test whether an image contains a cat by answering the following question: If the randomizer generates another random bitstring as inputs to

[88] The images in this figure are taken from the web (Forest, 2016; *Influenza in Cats* | *CDC*, 2024; Zielinski, 2020)

the UTM, what would be the probability that the UTM would generate the new image being tested (see Figure 16)?

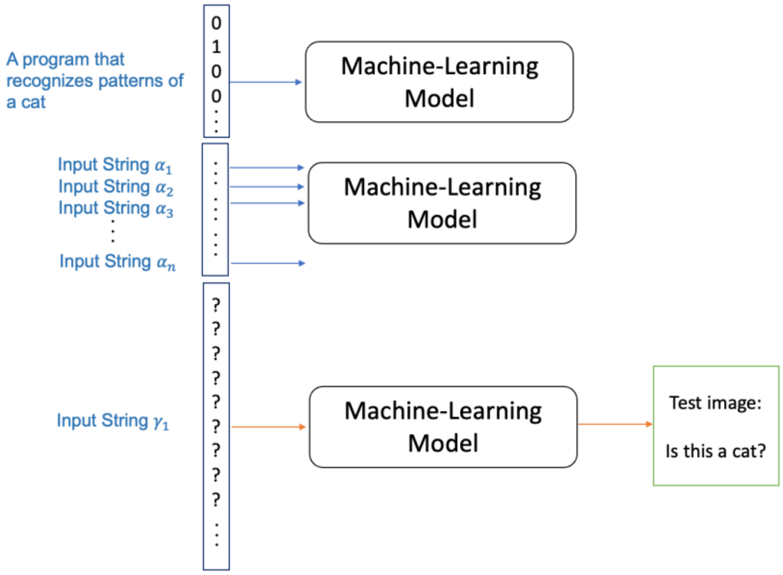


Figure 16: How to use Algorithmic Probability to test whether an image contains a cat.

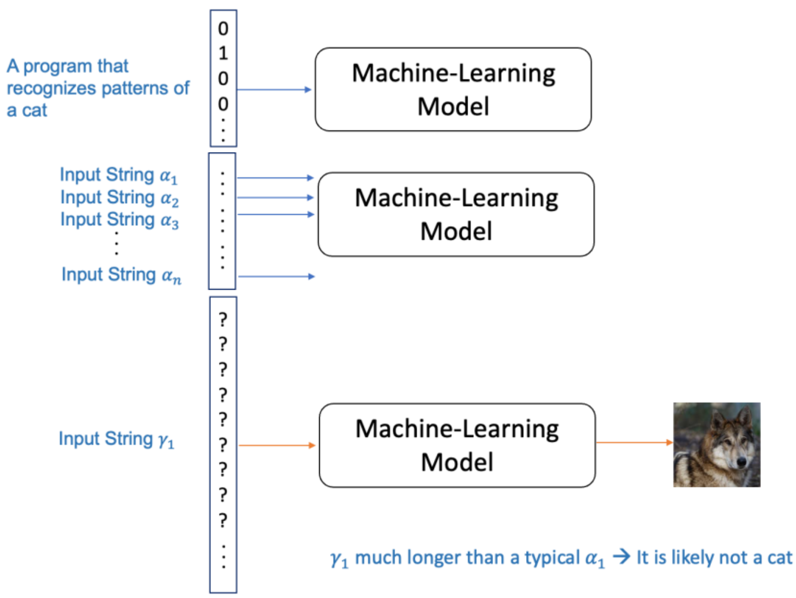


Figure 17: A longer input string is required to generate a dog image[89]

To see why, suppose we have an image of a dog and wants to test whether this is an image of a cat. Suppose further that the input strings α_1, α_2, α_3 ... α_i on Figure 15 have an average length of L_α. For the next output symbol to be the image being tested (the dog image), the algorithm comes up with an input string γ_1, which has a length of

[89] The image in this figure is taken from the web (Shirodkar, 2023).

L_{γ_1}. Not all patterns found in images of cats are shared by images of dogs, which also exhibit dog-specific attributes. Therefore, we would expect that $L_{\gamma_1} \gg L_\alpha$ because γ_1 needs the additional bits to describe dog-specific attributes (see Figure 17). The conditional probability that a randomized input would generate the dog image as the next symbol of an output sequence of cat images, given that we have come up with a minimal description on the sequence of cat images, would be $\left(1/2\right)^{L_{\gamma_1}}$. This conditional probability is much smaller relative to the conditional probability for a randomizer to generate a cat image, which should be relatively close to $\left(1/2\right)^{L_\alpha}$. Therefore, if we want to test whether or not a given image contains a cat, Algorithmic Probability can be used to compute the probability that a randomized input string would generate the given image, under the condition of having found a minimal description for a corpus of cat images as the output sequence to a UTM. We can then determine whether this image contains a cat based on whether this computed probability is close to $\left(1/2\right)^{L_\alpha}$.

5.6. The Implication of Algorithmic Probability on Deep Learning

According to Solomonoff, Algorithm Probability has been applied to genetic algorithms (2011, p. 5) and other optimization problems (1997, p. 17). But most significantly, it “can serve as a kind of ‘Gold Standard’ for induction systems” (1997, p. 22). That means, if time is not our concern and if computational resources are unlimited, this algorithmic model can find close to the best hypotheses for a body of data by collecting more data and by enlarging the scope of possible explanations (e.g., by increasing the time limit of computation for each hypothesis). In the following, I will explain why the argument for “Gold Standard” of Algorithmic Probability is equally applicable to a deep neural network, thus anticipating the eventual breakthrough in big data and deep learning.

In his Coursera lectures on machine learning, Andrew Ng explains how more data can drive the performance of machine-learning models.[90] The idea is to first increase the complexity of a model. This can be done by increasing the number of

[90] See Lecture 17 of Andrew Ng’s Coursera lectures on machine learning (Ng, 2017)

polynomial terms in a linear or logistic regression model, or by increasing the number of layers and nodes in a neural network[91]. In fact, deep learning is simply the training of a super complex neural network with lots of layers and nodes. Today, the number of hidden layers and nodes in a deep neural network may be in the thousands, millions, or even more. For instance, GPT-3 (GPT stands for "generative pre-trained transformers") [92], released by OpenAI in 2020, produced a gigantic model with 175 billion parameters (K.-F. Lee & Chen, 2021, p. 152). Google Brain, released one year later, is a language model with 1.75 trillion parameters (2021, p. 158). This number will only keep on increasing. Such complex models must be trained with a massive amount of data to make good predictions.[93] The more complex the model, the more data is required.

As I will try to explain, deep learning and big data were already anticipated in Solomonoff's Algorithmic Probability. Deep learning is a subfield of machine learning that focuses on training complex neural networks that are configured with many layers and nodes. The initial version of GPT model (GPT-1) has 12 layers, each containing 768 hidden units, and a total of 117 million parameters (Radford et al., 2018). GPT-2 has 48 layers, 1536 hidden nodes, and a total of 1.5 billion parameters (Radford et al., 2019). The latest version, GPT-3, has 175 billion parameters (T. B. Brown et al., 2020). The origin of artificial neural network can be traced to the Pitts-McCulloch artificial neuron/perceptron, which is a mathematical model that imitates the function of a biological neuron.[94] Warren S. McCulloch came up with the idea of artificial neuron and Walter Pitts "got it in proper form" by specifying an abstract schematics of an "impoverished" neuron in proper mathematical notations (McCulloch, 1974/2004, p.

---

[91] See Section 5.2 for the explanation on why the complexity of neural network is determined by the number of network nodes and layers.

[92] GPT is the engine behind the ChatGPT that is trending in popularity.

[93] Managing the bias-variance trade-off is a well-known diagnostic for debugging and tuning the performance of a learning algorithm. This diagnostic also underlies the rationale of big data analysis. For a low-bias, high-variance model, collecting more data for training can drive down the variance because more non-skewed training data allows a wider coverage of all possible patterns. If you have a massive training set with billions of data examples that basically exhaust all possible varieties, fitting a complex learning algorithm to this massive data set could also generalize well and make good predictions. The caveat is, you must choose a set of features that has sufficient information to predict accurately, such as recognizing whether or not a picture contains a car. Thus, if massive datasets are available, a typical strategy is to first choose a complex learning algorithm.

[94] See Section 5.2 for a brief introduction on how an artificial neuron imitates the biological model of a neuron.

353). By mimicking the functionality of a biological neuron, a perceptron with properly assigned weight, or a combination of these perceptions, can function as different logic gates (AND gates, OR gates, NOT gates, etc.).[95] As a Turing machine can be built with logic gates connecting to one another,[96] it can also be built as a network of artificial neurons. In “A Logical Calculus of the Ideas Immanent in Nervous Activity” (1943), McCulloch and Pitts came up with the mathematical proof that “a net made of threshold devices, formal neurons, can compute those and only those numbers that a Turing machine can compute with a finite tape” (McCulloch, 1974/2004, p. 353). Hence it is mathematically proven that any Turing machine can be implemented as a neural network.

Nonetheless, questions remain whether it is possible to train such a complex neural network with data. When Marvin Minsky and Seymour Papert wrote the book *Perceptrons* (1969/1972), it was generally believed at the time that no learning algorithm for multilayer network was possible (Hertz et al., 1999, p. 7; Sejnowski, 2018, p. 79). Their book demonstrates the limitation of a single layer neural network, raising doubts about the practicality of research into neural network. As a result, most of the computer science community left the neural network paradigm for almost twenty years (Hertz et al., 1999, p. 7). But the tide began to turn during the mid-1980s. David Ackley, Gregory Hinton, and Terrence Sejnowski invented a learning algorithm for multilayer network, which is described in their paper “A Learning Algorithm for Boltzmann Machines” (1985). Shortly after, Hinton, along with David Rumelhart and R. J. Williams, introduced the backpropagation learning algorithm for training multilayer neural network in “Learning Internal Representations by Error-Propagation” (1986). Backpropagation is now the method of choice for training deep neural networks. More recent technical literature has explored the question of Turing completeness of various neural network architectures. As Jorge Pérez, Javier Marinkovi, and Pablo Barceló explain in “On the Turing Completeness of Modern Neural Network Architectures” (2019), a “key requirement” for any “neural network architectures capable of learning algorithms from examples … is to

[95] See Lecture 8 of Andrew Ng’s Coursera lectures on machine learning (Ng, 2017)

[96] See Michael Sipser’s *Introduction to the theory of computation* (2013): “We can also simulate theoretical models, such as Turing machines, with the theoretical counterpart to digital circuits, called *Boolean circuits*,” which is “a collection of gates and inputs connected by wires. Cycles aren’t permitted. Gates take three forms: AND gates, OR gates, and NOT gates …” (Sipser, 2013, pp. 379–380)

have the capacity of implementing arbitrary algorithms, that is, to be Turing complete." Forms of neural network architectures proven to be Turing complete includes recurrent neural network (Siegelmann & Sontag, 1992) and the transformer model (Pérez et al., 2019).

Recall that the UTM in Algorithmic Probability can be substituted by any universal computer. Since a neural network can compute any function like a Turing machine, it can play the role of the Turing machine in Algorithmic Probability.[97] Increasing the complexity of a neural network is effectively the same as relaxing the time limit in the algorithm of Algorithmic Probability. When we increase the number of layers and nodes in an artificial neural network, we are essentially expanding the number of hypotheses under consideration. And in the same way that more data would narrow down the possible hypotheses in Algorithmic Probability, massive amount of data is required to tune the billion or trillion of parameters in a deep neural network to discard false hypotheses. The versioning of GPT exemplifies how deep learning may progress toward the completeness of Algorithmic Probability. For every new version, OpenAI would collect additional data to train GPT and design a larger neural network architecture to cover as many hypotheses as possible, given the size of the training data and the available computational power. Collecting more data, increasing the complexity of neural networks, and fabricating more powerful processor chips could result in deep learning models that approach the completeness of Algorithmic Probability.[98]

In Chapter 4, I brought up Turing's foresight that machine reasoning of "unorganized machines" could escape human understanding. Solomonoff's proofs also anticipates the opacity of complex machine-learning models in deep learning. Because his induction methods cover all possible hypotheses, "the proposed model is optimum with respect to all other conceivable models, many of which have not yet been

---

[97] By framing the neural network as a UTM in Algorithmic Probability, the design of a neural network is no longer concerned with replicating the actual physiological understanding of how the brain works, which AI and machine learning have no interest in.

[98] The "big data" theory has presumably been confirmed empirically in a study at Microsoft by Michele Banko and Eric Brill (2001). They examine the relationship between the choice of machine learning algorithms and the training set sizes, and the result shows that all algorithms, each of which can be made arbitrary complex, would yield similar performances. This result led to Andrew Ng's claim in his lecture: "it's not who has the best algorithm that wins. It's who has the most data." Nevertheless, an artificial neural network can be universal like a computer whereas other models cannot. So only neural network can be made to approach the completeness of Algorithmic Probability.

discovered" (1964, p. 17). Not only would his methods uncover new models, but they "are meant to bypass the explicit formulation of scientific laws, and use the data of the past directly to make inductive inferences about specific future events" (1964, p. 16). In other words, the optimal models are directly derived from the data themselves without introducing scientific knowledge or human reasoning. The data-driven models are sequences of bits that turn out to be the best software programs for generating the output sequence of symbols, and these software programs may be inexplicable in human reasoning. The opacity of these random bits anticipates the opacity of those billions and trillions of nodes in a deep neural network[99], as both encapsulate a machine-level reasoning beyond human-scale understanding.

The significance of Solomonoff's proof to machine learning parallels the Church-Turing thesis discussed earlier in Section 3.4. Just as the Church-Turing thesis gives an extreme outer limit of what it is possible to compute, Solomonoff's Algorithmic Probability gives an extreme outer limit on the subtle regularities that machine learning can identify in a large data collection. As I have argued, Algorithmic Probability is based on a machine-learning model that can always be improved upon by expanding the consideration of possible hypotheses and by eliminating invalid hypotheses with more data. This design scheme has been adopted in deep learning and big data, and the proof of "completeness" can be appropriated to explain the prowess of deep learning. In addition, Algorithmic Probability exhibits an opacity of machine-level reasoning that also characterizes deep learning. Both are associated with machine-level reasoning that can surpass human-scale reasoning for identifying subtle patterns in a body of data. With the way Algorithmic Probability anticipates deep learning, Solomonoff deserves the recognition of being "a pioneer scientist ahead of his time whose work was worthy of more than a Turing Award" (Dowe, 2013, p. 2).

---

[99] We can think of the hidden nodes in a neural network as sub-features that contribute to the predictive probability. Training a neural network would generate weights or parameters associated with each node over many iterations of back propagations. In the case of a neural network with only a hundred nodes, it is somewhat possible to confer meanings to these weights and nodes. Such a small neural network can be trained to recognize handwritten digits from one to ten, and a data scientist may inspect the weight of each node and try to guess how certain nodes may be associated with certain pen strokes, for instance, a straight line down here, a slant there (see Burrell, 2016: 6–7). But these are, at best, just educated guesses. When there are trillion of nodes, it is not possible for a human to explain the reasoning behind the model's predictions. Yet, these predictions could be better than any mechanism built based by a human.

5.7. Conclusion

This chapter explains the algorithm behind Solomonoff's Algorithmic Probability and why his proofs serve as the theoretical justification for machine learning and deep learning. As I have argued, early AI pioneers were inspired by Turing's papers on machine intelligence, and they justified their vision of AI by referring to the proof of universality in Turing machines. In the latest breakthrough in deep learning, the predictions of AI experts have their basis in the theoretical works by Solomonoff. In this chapter, I have outlined an explanation of Solomonoff's model of Algorithmic Probability, which encompasses the basic principles of deep learning and big data. These principles are: (1) make the model as complex as possible in order to cover a vast number of possible hypotheses, (2) train this model with massive amount of data to eliminate invalid hypotheses, and (3) train the model directly from data and acknowledge the opacity of machine-level reasoning.

In other words, the works that Dreyfus discarded (see Section 1.3) have turned out to be foundational for the breakthrough in deep learning.[100] Solomonoff's Algorithmic Probability and his proof of universal inductive inference support Kai-Fu Lee's claim that deep learning can recognize subtle patterns in data better than humans can. But this claim ought to be stated more concisely as follows: for any data pattern, there exists an algorithm that can recognize the pattern given no limit on computational power and training time. Realistically, there is always a limit to the available computational power and training time. Under such constraints, which change all the time, the job of AI practitioners is to explore what practical applications can be implemented, identifying new ways to design how to represent a real-world problem as numerical inputs and expected outputs to a machine-learning model. In other words, while there is the potential for deep learning to recognize any subtle pattern, there is no guarantee that it can recognize specific patterns in a given dataset under existing constraints in processing power and performance expectation. Such a project demands the continual experimentations of AI practitioners to produce workable AI models for any given constraints. Nonetheless, Solomonoff's proofs provide a rational ground for hope in such

[100] We ought not fault Dreyfus for missing out on the significance of Solomonoff's works though, since machine learning, and in particular, neural network, is disregarded as irrelevant until recently.

a research direction, much like the significance of the Church-Turing thesis for imagining what a computer can do.

At the same time, as long as AI research is exploring the potential of machine learning or deep learning, any claim about AI beyond pattern recognition capability should be questioned and challenged. This potential for deep learning to recognize any subtle data patterns demarcates any realistic projection of future development in AI. Qiufan Chen's storytelling in *AI 2041* imaginatively illustrate how our world may be transformed by such projections of future AI. The examples in *AI 2041*, along with recent AI inventions such as ChatGPT, demonstrate the vast potentials in applying the pattern recognition capability of deep learning. They also beg the question on the limit of AI, given that it is capable of learning from any data patterns. In fact, long before Ray Kurzwell wrote on the idea of Singularity (2005), Solomonoff already discussed the possibility of "a machine with a capacity many times that of the computer science community" as the final milestone of AI (1985). For him, if AI is capable of learning from all kinds of patterns, there seems to be virtually no limit to its intellectual capacity. Nonetheless, it is also difficult to fathom how AI can imitate the contingent development of knowledge and personality in human history. As mentioned in Chapter 1, Kai-Fu Lee points to three areas—creativity, empathy, and dexterity—where AI falls short of today. Yet some AI enthusiasts may still protest, what is creativity? Is creativity not simply randomness inserted into the regeneration of patterns that are deemed acceptable, such as the patterns in musical forms or in painting?

Such questions on the limitation of AI's potential demands a philosophical critique, which will be the subject of the following chapters. In the next chapter, I will turn to the philosophy of Gilbert Simondon, which conceives of a reality beyond the realm of digital algorithms. His works were formulated as a response to the philosophy of openness in cybernetics as advocated by Wiener, Wiener discusses his philosophy in *I Am a Mathematician, The Later Life of a Prodigy* (1956/1989). The autobiography describes his discomfort with the complete and closed system of logic in modern science, and his penchant to comprehend the disordered dynamicity between complex environments. Over the past three chapters, I have complicated the relationship between early developments in computation and the cybernetics movement.[101] I believe it is

[101] It is noteworthy that Turing himself never participated in the Macy Cybernetics Conferences.

better to narrate them as two distinct movements that mutually influence one another. It is true that on several occasions, the history of computing is marked by innovations that appropriate ideas from the intellectual current of cybernetics, as evident in the conceptualization of human-computer symbiosis, cyberspace, or machine learning. Nevertheless, decoupling the two intertwining movements can lead to the realization that the philosophy of cybernetics is not necessarily tied to computing or to the analytic philosophy that grounds the historical development of computing. By conceptually separating computer from cybernetics, we can gain clarity on the target of the critique of universality. The target of the critique ought not be the cybernetics movement, which is an interdisciplinary mosaic filled with mutually conflicting assumptions, united loosely by the theme of feedbacks, homeostasis, and the blurring of human-machine boundary. The target ought to be the universality in Turing's proof of the UTM, in Solomonoff's proof on Algorithmic Probability, and in Carnap's representation of the universe as a long sequence of symbols. The decoupling of cybernetics and computing allows us to appreciate the philosophy of openness behind cybernetics. As I will contend next, Simondon took this philosophy of cybernetics as the point of departure for his philosophical formulation, which led to his theories of individuation, concretization, and transduction.

Chapter 6.

Cybernetics and the Philosophy of Simondon

Me: Is there any creativity in DeepBach?

ChatGPT: The question of whether there is creativity in DeepBach, or any other computer program, is a complex and debated topic in the field of artificial intelligence and computational creativity.

From one perspective, creativity is often considered a uniquely human trait that involves novel and valuable outputs that are generated through a combination of knowledge, intuition, and imagination. Some argue that it is not possible for a machine to be truly creative because it lacks the subjective experience and consciousness that humans possess.

However, others argue that creativity can be defined more broadly as the ability to generate novel and useful outputs that are valued by humans. From this perspective, a machine such as DeepBach could be considered creative if it produces music that is novel, high-quality, and aesthetically pleasing to humans.

In the case of DeepBach, it is designed to generate new music that is similar in style to Bach's chorales, but is not a direct copy of existing pieces. The model has been shown to produce new and interesting variations of Bach's music that are stylistically consistent and can be considered musically coherent by experts. Therefore, some researchers argue that DeepBach is an example of computational creativity.

6.1. Overview

In the preceding chapter, I presented Solomonoff's Algorithmic Probability as an algorithm capable of discovering any subtle patterns in a large amount of data, anticipating the eventual breakthrough in big data and deep learning. A generic machine-learning algorithm, like Solomonoff's algorithm, would search through a large space of candidate programs and select the program that generates the data patterns shared by its training data. While this program can be applied to solve many pattern recognition problems, an even more straightforward application is the re-generation of data patterns in new forms. Thus, a machine-learning model trained to recognize the patterns in Bach's musical style is also a program that can re-generate new compositions imitating Bach's musical style. Indeed, among all the applications of deep learning, the production of artistic works using generative AI conjures up the impression that AI is capable of genuinely creative production, from graphics design to commercial music to even authoring a book[102]. This prompts the question: Would AI, empowered by deep learning, be capable of being genuinely creative? At this moment, the so-called creativity of generative AI is simply randomized regeneration of certain patterns. But if we operationalize human creativity as the ability to draw connections between background information that appears unrelated, creativity would then be a special case of the fringe consciousness that serves as the basis of Dreyfus' critique of symbolic AI (see Section 1.4). With deep learning capable of automatic language translation, which Dreyfus used as an example of fringe consciousness, would it also be possible to train an AI model with patterns of creativity such that AI becomes capable of creatively discovering relations from different aspects of life? To deliberate on this question about computational creativity, I now turn to the works of Gilbert Simondon, who formulated his philosophy of openness from Bergson's *Creative Evolution* (1922) and from the open character of cybernetics.

Simondon was a relatively inconspicuous French philosopher during his times, and his works have only begun to garner recognition among scholars in recent decades (Feenberg, 2017a, p. 66). Plenty of commentaries have since come out, explaining his philosophical concepts while acknowledging their significance as a critical response to cybernetics. They mostly give the impression that Simondon is critical of cybernetics.[103]

---

[102] E.g., ChatGPT is the author of a book in recently published in Korea (Hwang, 2023).

[103] Cecile Malaspina also points out that "many commentators will be quick to point out that Simondon is critical of cybernetics and information theory" (2019). Here are some examples. In "Governing progress: From cybernetic homeostasis to Simondon's politics of metastability" (2022), Andrew Bardin writes, "[f]ollowing Gilbert Simondon, we take the cybernetic notion of dynamic stability ('homeostasis') as paradigmatic of the hyper-modern condition" and "[t]he connection we establish between cybernetics and neoliberalism will eventually allow us to use Simondon's theory against both." In "Being with Technique–Technique as being-with: The technological communities of Gilbert Simondon" (2019), Susanna Lindberg writes, "Simondon thinks that Norbert Wiener's cybernetics is too limited a theory because it examines automats [*sic*] too exclusively as if they were entirely closed systems." In *Gilbert Simondon and the Philosophy of the Transindividual* (2013), Muriel Combes emphasizes Simondon's resistance to cybernetics: "Simondon does indeed use the term "feedback," but because of his resistance to cybernetics, and because I see something very different at work in his philosophy than in theories of autopoiesis" (2013, p. 118).

Nevertheless, according to Pascal Chabot, Simondon was "[i]nspired as much by Ionian physiology as by cybernetics" (2013, p. 1).[104] He was in fact

> among the first to bring cybernetics to France. He read Wiener's writings as soon as they were published. Simondon shared Wiener's enthusiasm for a transdisciplinary theory organized around mutually agreed-upon concepts. He described phenomena using operational terms and adopted the vocabulary of the cyberneticists, with recurring references to communication, control, relations, functions, actions and reactions. (2013, p. 53)

As I will explain in this chapter, if we read Simondon's works in parallel with Wiener's *Cybernetics* (1948/2007) or with the transactions of the Macy Cybernetics Conferences (Pias, 2003a), it becomes apparent that Simondon was not only critical of cybernetics but also appreciative of key ideas in cybernetics, to the extent that he incorporated and appropriated these ideas in his philosophical inquiry.[105] Many key philosophical concepts in Simondon can be viewed as derivatives of ideas from cybernetics. For instance, it is possible to identify associations between recurrent causality and negative feedback, between concretization and homeostasis, between individuation and ontogenesis or phylogenesis, and between transduction and the communication of chemicals between biological entities.[106]

Hence, while Simondon and Heidegger both offered critiques of cybernetics, the critique of Simondon has a different character from that of Heidegger, who abhorred the

---

[104] Yuk Hui makes a similar point: "The richness of the discussions in cybernetics are rarely taken into account of understanding Simondon, the Simondonian scholars seem to emphasize how different is Simondon from the cyberneticians without looking into the legacy" (2011).

[105] Andrew Iliadis makes aware the importance to engage with the sciences and cybernetics to interpret Simondon: "What Deleuze did not point out, and what many English readers of Simondon have heretofore failed to pick up on, is that in articulating this new philosophy Simondon was simultaneously engaged in conversation with some of the most technically advanced scientists, engineers, and mathematicians of the twentieth century. Any real understanding of Simondon's approach to individuation – most central of all Simondonian concepts – must acknowledge the privileged position that Simondon gave to notions from within engineering, physics, and especially cybernetics in his original ontology" (2013, p. 1). Note that, whereas Iliadis engages in a discussion of information ontology in Simondon's concept of individuation, this chapter and the next are concerned with other theories in cybernetics and in modern physics.

[106] Note that Chabot never further identifies Simondon's associations with cybernetics at this level of details. He does note that Simondon celebrates the new relationship between people and machines invented by cybernetics: "Cybernetics invented a new relationship between people and machines. Simondon celebrates this. He demonstrates that information technologies represent a real step forward: they allow for the successful coupling of human and machine" (2013, p. 71). We will come back to this relationship in Chapter 6.

totalizing and universal implication of cybernetics.[107] As I will further argue, Simondon's writing captures the ethos of cybernetics: the bringing together of knowledge domains with contradictory assumptions, and the resulting sociotechnical innovations that transcend the inherent contradictions and incompatibilities. According to Simondon in "Cybernétique et philosophie" (1953/2016a) and "Épistémologie de la cybernétique" (1953/2016b), people often confuse information theory with cybernetics (1953/2016a, para. 17).[108] In fact, information theory is only a branch of cybernetics (1953/2016b, para. 7).[109] What initiated the cybernetics movement was this recurrence of the effects of activity on activity, which is called reaction, feedback, or internal resonance (1953/2016a, para. 17).[110] In contrast to information theory, the philosophy behind this notion of feedback in the cybernetics movement, as outlined by Norbert Wiener, is suggestive of a possible trajectory of co-evolution between the human, the social, and technology. Simondon further develops this co-evolution in his philosophy of individuation.

Therefore, in contrast to Heidegger's critique, Simondon is only partially critical of cybernetics, particularly on the reduction of communication into one of digital information and on the mechanization of the living beings that fits the symmetric approach to life and machines in cybernetics. His critical reflection of cybernetics actually spawns out of his philosophical formulation that is primarily inspired by cybernetics, along with all the breakthroughs in scientific discoveries leading up to his time. This association with cybernetics becomes even clearer when we examine Wiener's reflections on his exposure to philosophy that eventually led to his exploration of cybernetics. This philosophy embraces the themes of openness and complexity in opposition to the closed

[107] See Chapter 2 on the discussion on Heidegger's critique of cybernetics.

[108] "Cette psychanalyse défensive, qui fait le bilan du mythe et de la réalité dans la cybernétique pour la réduire à une théorie de l'information, confond l'amour populaire pour le robot et la technologie des systèmes holiques. Un robot n'a rien de cybernétique" (1953/2016a, para. 17).

[109] "Toutes les définitions qui ont été données jusqu'à ce jour de la Cybernétique désignent les cybernétiques particulières, réciproques de telle ou telle science structurale, plutôt que la Cybernétique universelle. L'étude de la quantité d'information est une branche de la Cybernétique, comme l'étude des mécanismes téléologiques, ou encore celle des relations d'asservissement et de commande" (1953/2016b, para. 7).

[110] "Cette récurrence des effets de l'activité sur l'activité se nomme réaction, feed-back, ou résonance interne. Avec la réaction commence le système cybernétique" (1953/2016a, para. 17).

system of traditional sciences, and such themes were later taken up by Simondon in his philosophical inquiry.

In the following, I will summarize Wiener's philosophy of cybernetics and its emphasis on the themes of openness and complexity. I will then introduce Simondon's theories of individuation and concretization while pointing out how these theories are associated with cybernetics. Two key concepts that Simondon adapts from cybernetics are "recurrence of causality" and "transduction." Somewhat ironically, these concepts, derived from cybernetics, also form the basis for Simondon's critical responses to cybernetics' mechanization of the living and its blurring of boundary between life and machine. I will conclude by revisiting the question of whether AI can be creative and exhibit life-like quality in light of Simondon's critique.

6.2. Wiener's Philosophy of Cybernetics

Earlier in Chapter 3, I argued that the intellectual current of cybernetics is filled with contradictions and conflicts between disciplines. These incompatible disciplines were loosely held together by common themes such as negative feedbacks and homeostasis, whose interdisciplinary implications were summarized by Wiener in *Cybernetics* (1948/2007).[111] As I will explain below, behind these cybernetics ideas is an approach to scientific exploration that seeks to theorize irregularities and complexities that escape the closed system of traditional scientific method.

Originally trained as a philosopher before becoming a mathematician, Wiener reflected on the philosophy of cybernetics in his autobiography *Norbert Wiener—A Life in Cybernetics* (1956/2018): "The whole background of my ideas on cybernetics lies in the record of my earlier work" (1956/2018, p. 459). In his earlier work, Wiener repudiates the closed system of logic in modern scientific method and expresses respects for irregularities, complexities, and open systems:

> Because I had studied harmonic analysis and had been aware that the problem of continuous spectra drives us back on the consideration of

[111] According to David A. Mindell in *Between human and machine: feedback, control, and computing before cybernetics* (2002), Wiener's publication reflects more a summary of the discussion from the meetings he attended at the Macy Cybernetics Conference than his own ideas. Mindell's book also documents a long history of technical designs related feedback and homeostasis.

> functions and curves too irregular to belong to the classical repertory of analysis, I formed a new respect for the irregular and a new concept of the essential irregularity of the universe. Because I had worked in the closest possible way with physicists and engineers, I knew that our data can never be precise. …
>
> It is no coincidence that my first childish essay into philosophy, written when I was in high school and not yet eleven years old, was called "The Theory of Ignorance." Even at that time I was already struck with the impossibility of originating a perfectly tight theory with the aid of so loose a mechanism as the human mind. And when I studied with Bertrand Russell, I could not bring myself to believe in the existence of a closed set of postulates for all logic, leaving no room for any arbitrariness in the system defined by them. Here, without the justification of their superb technique, I foresaw something of the critique of Russell which was later to be carried out by Gödel and his followers, who have given real grounds for the denial of the existence of any single closed logic following in a closed and rigid way from a body of stated rules.
>
> To me, logic and learning and all mental activity have always been incomprehensible as a complete and closed picture and have been understandable only as a process by which man puts himself *en rapport* with his environment. (Wiener, 1956/2018, pp. 459–460)

Wiener raises the issue that "essential irregularity of the universe" escapes "the classical repertory of analysis." Scientific hypotheses stipulated in formal mathematical functions of classical analysis cannot account for irregularities. In addition, these hypotheses are formulated to account for data measured by physicists and engineers, and such data "can never be precise." This is essentially consistent with the arguments made in Thomas Kuhn's *The Structure of Scientific Revolution* (1962/1996) and later in science and technology studies (STS). The modern scientific project generalizes from observations of objects in a closed system. These objects have been extracted from their contexts and isolated from changes in their environment. The resulting abstract knowledge has little to say about the real and concrete world which is made up of complex relations between open environments. The science of closed systems, such as experimental and theoretical physics, does not take account "nature's overwhelming tendency to disorder" (Wiener, 1956/2018, p. 460), and "[t]hese were the ideas [he] wished to synthesize in [his] book on cybernetics" (1956/2018, p. 460). Rather than seeking a unified theory, the holy grail of modern scientific pursuits, Wiener took Gödel's incompleteness theorem as his point of departure. His aim was to bring new understanding about the disordered dynamicity of the interactions and communications between complex environments.[112] In other words, Wiener's cybernetics was an attempt to revolutionize modern sciences, turning them from the study of closed systems of orders into the study of open systems of disorders. By shifting the focus of scientific studies to open systems, in which there is no assumption of a fixed and controlled environment, Wiener derived the notions of negative feedbacks, which, as mentioned earlier, initiated the cybernetics movement according to Simondon.

Feedbacks as purposeful active behavior in a disordered, open environment characterize cybernetics. Up to this point, Simondon would agree with the general direction of cybernetics as the scientific study of complex phenomena. But whereas Simondon identified such complexity as the point of departure for his philosophy of openness, scientists such as McCulloch took the cybernetics movement into another direction. Cyberneticians believed that their field can overcome the philosophical dualism and reductionism in the modern epoch. The result would be the universalization across different fields. As McCulloch concludes in "The Beginning of Cybernetics" (1974/2004), cybernetics "is ready to officiate at the expiration of philosophical Dualism and Reductionalism … Our world is again one, and so are we" (1974/2004, p. 360). Relating the inner milieu with the outer milieu in a feedback loop, would resolve the subject-object dualism, and the focus of disordered and open environment would address the reductionism of the classical analysis in modern science. To cyberneticians like McCulloch, the new science of cybernetics has a deeper philosophical implication behind the technical prototypes and inventions: the implication of unifying a disjointed world into one world, between subject and object, between isolated knowledge disciplines, between humans and machines.

So, it is quite possible that Heidegger's statement, that cybernetics would take the place of philosophy,[113] originates from the scientists' opinion about the universality of cybernetics. But this speculation in fact takes a conceptual leap from the notion of feedback as a generic mechanism across boundaries and between milieux, to the belief that all beings can be universally mechanized as processes in feedback loops. Hence, as I have argued in Chapter 3, this belief about the universality of cybernetics could

---

[112] Note that references to Gödel's incompleteness theorem often come up in the transactions of the Macy conferences.

[113] See Chapter 2 on the discussion on Heidegger's critique of cybernetics.

never be carried out in practice due to the inherent contradictions between disciplinary knowledge. It is true that feedbacks can be found everywhere, but no knowledge discipline is reducible to feedback processes. With his breadth of knowledge in sciences and technologies, Simondon was likely aware of the untruth in the unfounded claim that cybernetics is a universal science. Therefore, unlike Heidegger who finds in cybernetics the materialization of the technological will due to its universality, Simondon's philosophy rebuts the very claims that lend support to the universality of cybernetics. This includes the claim that the human and the machine are both made up of processes of information and are therefore essentially the same. At the same time, Simondon was appreciative of the genuine insights behind the theorizing of irregularities and disorderly systems across multiple milieux in cybernetics. As I will explain in the remainder of this chapter, the philosophy of Simondon can be viewed as a further development of the thoughts on complexities that were preliminarily explored in cybernetics.

6.3. Cybernetics and Simondon's Philosophy of Individuation

Simondon's most prominent philosophical works are *L'individuation à la lumière des notions de forme et d'information* (1964/2005; hereafter *ILNFI*) and *Du mode d'existence des objets techniques* (1958/1989; hereafter *MEOT*). *ILNFI* is the main thesis of Simondon's doctoral dissertation. It was originally published in two parts, the first part on the ontogenesis and phylogenesis of physical and biological beings, the second part on the growth of individual psyches and the development of social collectives. *L'individu et sa génèse physico-biologique* was published in 1964, whereas *L'individuation psychique et collective* remained unpublished until 1989. *MEOT*, his earliest publication, is the complementary thesis, but was immediately turned into publication after his doctoral defence in 1958.[114] Building on Henri Bergson's *Creative Evolution* (1922), Simondon develops his philosophy of individuation in *ILNFI*, and extends this philosophy to his analysis of technology in *MEOT*, in which he provides a theory on sociotechnical innovation with his theory of concretization. His writing alludes to an eclectic source of knowledge and examples drawn from various fields in science

[114] On a side note, *ILNFI* did not appeal to a wide readership until it was brought up by Gilles Deleuze and then later by Bernard Stiegler in their works. In contrast, *MEOT* was immediately well-received by a broad audience after publication (Feenberg, 2017a, p. 66; "Gilbert Simondon," 2023).

and technology. This multidisciplinary discourse gives the impression that Simondon is responding to the discussions at the Cybernetics conferences. In this section, I will present an overview of the philosophy of individuation in *ILNFI* and allude to its conceptual connections with cybernetics. I will then discuss *MEOT* and his theory of concretization in the next section.

In *ILNFI*, the latest scientific discoveries in physics and biology serve as the material basis for problematizing the traditional notion of individual and individuation, leading to his theorizing on the perpetual evolution of new forms (individuation), which has its source of potentiality in conflicts and contradictions (pre-individuality). These theories come with the implicit assumption that classical philosophy was formulated without the awareness of phenomena observable today via more advanced scientific instruments. A prototypical example is the matter-form paradigm derived from the brick formation (Simondon, 1964/2020, pp. 21–54).[115] Initially, a brick maker would begin the process by filling a wooden mould with clay. Under the matter-form paradigm, clay is a malleable, formless substance, which requires the wooden mould to give it the form of a brick. Simondon argues that the wooden mould is not a pure form, but material that requires technical treatment to become hardened and to assume the appearance of a form. Neither is this clay purely indeterminate matter. Rather it is processed material with molecular properties that determine its porosity and density. When heated, compressed clay would expand and press up against the wooden mould, which acts as an opposing force to the expansion. Thus it is not the form of the mould, but this exchange of force and energy that produces the hardened form of clay.

This limitation of matter-form paradigm persists in our understanding not of only clay and brick, but also of “many events of formation, genesis, and composition in the living world and the psychical domain” (1964/2020, p. 21). These facts include the different forms that matter may take under different phases in crystallization (1964/2020, pp. 68–87), as well as the obfuscated boundaries of individuality in many physical and living beings, from quantum mechanics (1964/2020, pp. 135–148) to the society of bees and ants (1964/2020, p. 337) to the colonies of corals (1964/2020, pp. 180, 208).[116]

[115] More specifically, the target of the matter-form critique is the philosophical theory by Aristotle called hylemorphism.

[116] I will elaborate on the obscure zones of quantum theory and crystallization for the deliberation of pre-individuality in Chapter 7.

Each of these examples operates in some “obscure zone” outside the matter-form paradigm and problematizes the notion of a complete and coherent individual. The matter-form paradigm requires a “reduction of the entire spectrum of reality,” which includes these obscure zones, to “its extreme terms considered as matter and form” (1964/2020, p. 351). These obfuscated boundaries of individuality seems reminiscent of the cybernetic notion of boundary-crossing across milieux. But rather than deconstructing the boundary between the human and the machine,[117] Simondon identifies the “obscure zone” as the reservoir of potentiality prior to the formation of individuals, in such a way that individuals are inherently associated in relations to one another and to the environment that engender the individuation. Thus in Simondon’s philosophy, beings are inherently in relation.

Simondon emphasizes this inherent relationality in his philosophy of individuation, which was formulated to overcome the matter-form paradigm. The process of crystallization serves as a paradigmatic model for individuation and consequentially, for the relationality between individuals and milieux. A super-saturated solution contains more solute than the equilibrium solubility allows. It remains in an amorphous state until it is inseminated with a crystalline germ. Crystalline structures would then begin to form in extension of the crystalline germ. As long as the region of the newly constituted structure is in contact with the solution, the activity will continue to propagate. Analogous to the actual events in crystallization, an individuation undergoes

> a physical, biological, mental, or social operation through which an activity propagates incrementally within a domain by basing this propagation on a structuration of the domain operated from one region to another: each structural region serves as a principle and model, as an initiator for constituting the following region, such that a modification thereby extends progressively throughout this structuring operation. (Simondon, 1964/2020, p. 13)

Simondon uses the term “transduction” to denote the discharge of pre-individual potential energy that would bring about the appearance of form-taking (Garelli, 1964/2020, pp. xxii–xxiii). The potential energy of a supersaturated solution comes from the relations between fields of extreme tension, from the incompatibility between an overdose of solutes and the equilibrium solubility of the solution. What Simondon calls

---

[117] For instance, as Donna Haraway did in “A Cybernetic Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century” (1991).

the state of *disparation* refers to the incompatibility causing the extreme tension between two disparate realities. The crystalline germ, which is essentially a very small lattice structure[118], introduces a new dimension to this state of disparation, prompting ions to latch onto this lattice structure to achieve a more optimal energetic state. In other words, the small crystal structure introduces a new kind of order that resolves the extreme tension between solutes and equilibrium solubility. Simondon uses the analogy of vision to illustrate how introducing a third dimension brings about the resolution of disparation (1964/2020, pp. 15, 229–230). Two incommensurable two-dimensional planes are conjugated into the projection of a single three-dimensional space. Discrepancies between the two-dimensional spaces hold positive potentials for the visualization of a three-dimensional space. Therefore, apparently negative aspects have the potentiality to become positives, and a paradigm shift can actualize such potentiality. In this analogy, the lattice structure of a crystal corresponds to the third dimension that brings resolution to the incompatibilities between the disparate regimes of supersaturation and equilibrium solubility.

On the surface, the crystalline example is a prototype for understanding ontogenesis. Describing ontogenesis as mechanical processes appears to be one of the main branches in the cybernetics movement. As discussed in Section 4.3, when cyberneticians deliberated on the meaning of "learning" in the context of machines, they also brought up the ontogenetic evolution of intelligence in organisms. Nevertheless, since the individuation of crystallization involves the operation of transduction that actualizes potentiality, the example of crystallization is as much about ontogenesis as it is about the potentiality of pre-individual reality:

> individuation appears, on the one hand, as ontogenesis and, on the other hand, as an operation of a pre-individual reality that not only produces the individual, the model of substance, but also produces the energy or the field associated with the individual; only the associated field-individual pairing accounts for the level of pre-individual reality. (1964/2020, p. 160)

Rather than mechanizing ontogenesis, Simondon attributes the growth of an organism to the potentiality originating from a reality prior to the forming of the individual. In the case

---

[118] See Section 7.3 for an elaboration about the atomic and molecular activities involved in crystallization.

of crystallization, this “pre-individual” reality is the metastable crystalline solution.[119] In general, the “pre-individual reality” is the primordial unity that has undertaken a phase-shift into an individual and an associated field of potential energy. Until the potential energy is exhausted, the individual can continue to individuate, to transition to another phase. For instance, the liquid phase continually transitions to the solid phase during the process of crystallization. The field of potential energy is filled with unresolved tension, and individuation is the continual phase shifting of an individual to resolve the tension in a preceding phase. These tensions remain unresolved until the exhaustion of the potential energy.

Transduction and pre-individuality are the basis for Simondon’s theory on the relationship between different levels of individuations:

> Logically, [transduction] can be used as the basis of a new type of analogical paradigmatism in order to pass from physical individuation to organic individuation, from organic individuation to psychical individuation, and from psychical individuation to the subjective and objective transindividual, all of which defines the plan of this research. (Simondon, 1964/2020, p. 14)

Accordingly, the metastable milieu for organic individuation carries a remnant of pre-individual potentiality from physical individuation, and this potentiality is transduced into a process of organization and integration, which engenders organic individuals that can be differentiated from other individuals. This remnant of pre-individual potentiality from physical individuation is transduced to bring about organic individuation, and the same type of operation transduces organic individuation to psychical individuation, and psychical individuation to collective individuation, in which “the collective … is a transindividual reality” (1964/2020, p. 179). Simondon uses the term “transindividual” to denote “across beings” (1964/2020, p. 344), and the rapport of beings in transindividual reality is the basis behind human language and communication (1964/2020, p. 345).[120]

---

[119] Grasping the physics behind crystalline lattices can clarify Simondon’s concept of pre-individuality, which I will discuss in Chapter 7.

[120] A more thorough discussion on Simondon’s theory of transindividuality and its implication on signification and language is provided in Section 8.3.

6.4. From Cybernetics to Concretization and Transduction

While the philosophy of individuation is Simondon's main focus of *ILNFI*, his theory of concretization for re-establishing the rapport between technics and the natural or social milieu is the key formulation of *MEOT*.[121] In Chapter 3, I explained how the cybernetic notions of feedbacks and homeostasis are abstract schemes for mechanizing relations between heterogeneous milieux. Simondon also worked on characterizing the relation between heterogeneous domains, but he wanted to do so without succumbing to the cybernetic project of mechanization and automation. This effort led to his formulation of concretization.

In the most straightforward sense, concretization is a theory of how technical objects evolve, of how elegant and inventive designs can eliminate redundant functions and reduce complexities. Technical design is often a matter of attenuating complexity because upgrading a technical system would easily lead to an increase of complexity, eventually high enough to cause system breakdowns.[122] Efforts to enhance a complex technical object may come to a dead end unless its modules and their relations are redesigned to become simple enough for further expansion and development. Thus, when a technical object evolves, its new design needs to become more "concrete" than older generations, with a reduction in complexities. As a technical system becomes concrete, it is "tending toward internal coherence, toward a closure of the system of causes and effects that exert themselves in a circular fashion within its bounds" (1958/2016, p. 49). In a deeper philosophical sense, Simondon's formulation of concretization in *MEOT* is derived from his analysis on the ontogenesis and homeostasis of the living in *ILNFI*. Analogous to how heterogeneous milieux tend toward equilibrium in homeostasis, a concretizing technical system tends toward internal coherence between multiple technical or natural milieux. This internal coherence comes from the "closure of the system of causes and effects that exert themselves in a circular fashion" (1958/2016, p. 49). Simondon calls this closure "recurrent causality" or "recurrence of causality," a term that he uses in both *ILNFI* and *MEOT*.

---

[121] As I will discuss in Chapter 7, Simondon puts forth a figure-and-ground paradigm, in which technics are the figures and the natural or social milieu is the ground.

[122] This is what Simondon means when he talks about pushing a technical evolution toward a "fatal hypertely" (1958/2016, p. 58).

In *ILNFI*, recurrent causality takes on a general, non-technical meaning, indicating how two processes mutually and recurrently cause changes to each other. This recurrence is analogous to the notion of cybernetic feedbacks, as both establish a two-way communication channel between milieux. In fact, Simondon uses the term feedback as synonymous with the term "recurrent causality" in "Cybernétique et philosophie" (see 1953/2016a, para. 17). But whereas cybernetic feedbacks denote a recurrent flow of information, "recurrence of causality" refers to a recurrence of mutual causation between two or more milieux. This paves the way for Simondon's differentiation between the biological notion of homeostasis in the living beings and the cybernetic notion of homeostasis. This difference is explained in *ILNFI*: "homeostasis is related to external conditions of transduction due to which the being utilizes the equivalence in external conditions as safeguards for its own stability and its internal transduction" (1964/2020, p. 172). Whereas Ashby's homeostat tends toward equilibrium over feedbacks, homeostasis in the living being serves as both "safeguards for its own stability *and its internal transduction* [emphasis added]." In a transductive operation, the external milieu causes the transformation of the internal milieu, whose reactions in turn cause the external milieu to transform in a "recurrence of causality." Therefore, "recurrence of causality" denotes both a process of integration and a process of differentiation. It is a process of integration because the operation brings a coupling stability between milieux. It is a process of differentiation because the operation brings respective changes to the milieux. This is what Simondon means when he wrote, "[l]ife would therefore be conditioned by the recurrence of causality due to which a process of integration and a process of differentiation can receive a coupling while remaining distinct in their structures" (1964/2020, p. 173).

Concretization incorporates this understanding of recurrent causality in homeostasis. A concretizing technical system "tend[s] toward a closure of the system of causes and effects that exert themselves in a circular fashion within its bounds" (1958/2016, p. 49). Recurrent causality effects mutually structural changes between technical elements or between these elements and their associated milieu, which Simondon defines as a simultaneously technical and natural milieu. This associated milieu is "a milieu that the technical object creates itself and that conditions it, just as it is conditioned by it" (1958/2016, p. 59), in a recurrence of causality. But concretization also differs from homeostasis in the same way that "[t]he ontogenesis of the living being cannot be conceived from the notion of homeostasis alone, or by self-regulation of a perpetuated metastable equilibrium” (1964/2020, p. 204). According to Simondon’s analysis of ontogenesis in *ILNFI*, there is “an inner problematic of being” in the living being:

> [t]he state of a living being is like a problem to be solved, of which the individual becomes the solution through successive assemblages of structures and functions. The young individuated being could be considered as a system as a carrier of information, in the form of pairs of antithetical elements, linked by the precarious unity of the individuated being whose internal resonance creates a cohesion. (1964/2020, p. 204)

The development in ontogenesis “could then appear as the successive inventions of functions and structures that solve, step by step, the internal problem carried as a message by the individual” (1964/2020, p. 204). These “successive inventions of functions and structures that solve … the internal problem” in the living, as described in *ILNFI*, are appropriated to the technical realm by Simondon in *MEOT*. Concretization is the technical correspondence to ontogenesis of the living, as a technical system evolves by new inventions of functions and structures that resolve the disparations between system modules. Thus a concretizing technical system tends toward an internal coherence over successive inventions of technical “organs” that resolve the “inner problematic” of a technical being.

In *MEOT*, Simondon alludes to Jean-Claude Guimbal’s invention of a hydroelectric generator to illustrate the concretization of a technical system, the creation of an associated milieu, and the recurrence of causality. Prior to the Guimbal’s invention, it was unthinkable to insert an electric generator into a penstock, which is a conduit or pipe for conducting water in a hydroelectric power plant. A generator was too large to be lodged into the conduit of a penstock because it needs to be wrapped around by an outer layer that addresses the problems of electrical insulation and water-tightness (i.e., impermeability to water). Guimbal’s invention makes it possible to reduce to the size of a hydroelectric power plant. As Andrew Iliadis explains, “it is due to Guimbal’s genius in realizing that the [generator] could be cooled in water that this concretization could occur, since it is by virtue of the automatic water cooling that the [generator] could be built much smaller" (2015, p. 91).[123] According to Simondon, Guimbal invented a very

[123] The original quote from Iliadis’ article is: “it is due to Guimbal’s genius in realizing that the turbine could be cooled in water that this concretization could occur, since it is by virtue of the

small generator contained in a crankcase filled with pressurized oil. The oil serves to insulate the generator from electricity and prevents the seepage of water into the crankcase through the shaft packings because the pressure of the oil is greater than the pressure of the water outside the box. If this generator is put in open air, it would be destroyed due to overheating. It is only when the generator is placed in the conduit with the water turbine that the problem of overheating is resolved, as the oil in the generator would transfer the generated heat from its winding to its crankcase where the heat would dissipate by the cooling of the water. Thus "the insertion of the generator into the conduit *renders itself possible* by simultaneously authorizing the energetic cooling by water" (Simondon, 1958/2016, p. 57 emphasis in original).

This design illustrates the recurrent causality between technical objects and the very condition that makes possible their existence. Or as Yuk Hui puts it, a technical object in recurrent causality contributes to the condition that makes possible its existence: "it is situated in a system and in reciprocal relations with other parts; it adapts itself to the system while at the same time modifying the system, which in turn conditions its further mode of operation; it becomes its own condition through the feedback of the whole organic system" (2019, sec. 32). Over the recursive relations between technical objects and their technical milieu, concretization is "conditioned by an invention that *presupposes the problem to be resolved*" (Simondon, 1958/2016, p. 57 emphasis in original). It is

> a process that conditions the birth of a milieu rather than being conditioned by an already given milieu; it is conditioned by a milieu that only exists virtually before invention; there is invention because there is a leap that takes place and is justified by means of the relation that it brings about within the milieu that it creates: the condition of possibility of this turbo-generator couple is its realization; … One could say that a concretizing invention realizes a techno-geographic milieu (in this case the oil and water in turbulence), which in turn is a condition of possibility of the technical object's functioning. *The technical object is thus its own condition, as a condition of existence of this mixed milieu* which is simultaneously both technical and geographical. (1958/2016, p. 58 emphasis in original)

If an engineer experiments with every imaginable technique to reduce the size of an electric generator as a pre-condition for putting the generator into a penstock, she or he

---

automatic water cooling that the turbine could be built much smaller" (2015, p. 91). But it is actually the generator, not the turbine, that could be cooled in water and could be built much smaller.

would never arrive at the elegant invention of using the condition inside the penstock to resolve the sizing problem. Therefore, straight-line thinking, either deductively or inductively, cannot lead to the inventive realization of simultaneous mutual conditioning between technical elements and a technical-geographical milieu that is yet to exist. Every invention of a technical object involves a transductive realization of a new techno-geographical milieu in which technical elements are put in relations with each other and with physical entities from nature. To Simondon, "[t]he only technical objects that can be said to have been invented, strictly speaking, are those that require an associated milieu in order to be viable" (1958/2016, p. 59).

The recurrent causality in concretization resolves the "inner problematic" in a transductive operation, bringing the contradictions in one frame of reference into a harmonious relation using another frame of reference. This inventive character of transduction, fully developed in the formulation of concretization in *MEOT*, has already been anticipated in *ILNFI*: "In the domain of knowledge, [transduction] defines the veritable measure of invention, which is neither inductive nor deductive, but transductive, i.e., corresponds to a discovery of the dimensions according to which a problematic can be defined" (1964/2020, p. 14). A genuine invention is transductive, discovering new dimensions that transduce contradictory tensions into mutually supporting structures in a recurrence of causality. Contrary to induction,

> transduction is a discovery of dimensions whose system makes the dimensions of each of the terms communicate, such that the complete reality of each of the terms of the domain can become organized into newly discovered structures without loss or reduction; resolving transduction *operates the inversion of the negative into the positive* … (1964/2020, p. 15 emphasis in original)

This transduction here, as defined in ILNFI, is generic for physical, biological, mental and social operations. But if we read this definition in the context of technical knowledge and invention, it precisely maps his definition of concretization in *MEOT*. In fact, the concretizing character of transduction is evident in the following passage in *ILNFI*:

> [T]ransduction is characterized by the fact that the result of this operation is a concrete fabric including all the initial terms; the resulting system is made of that which has become concrete and includes the whole concrete; the transductive order conserves the concrete and is characterized by the *conservation of information*, whereas induction requires a loss of information … (1964/2020, p. 15 emphasis in original)

Transduction is an operation that would result in a "system … which has become concrete." In other words, concretization is simply transduction recontextualized in the domain of technical knowledge. In concretization, the "initial terms" would correspond to the scientific representations based on which technical schema can be conceptualized, and the concretizing system is the site for the actual behavior of the technical elements towards each other and with the associated milieu. The coming together of all the "initial terms" in a single move would result in a system that has become concrete. In *MEOT*, we can also find passages on concretization that are reminiscent of the explanation of transduction in *ILNFI*. Simondon explains in *MEOT*, "[c]oncretization gives the technical object an intermediate place between the natural object and scientific representation" (1958/2016, p. 49). The "intermediate place" represents the gap between abstract scientific representations and a technical object that tends toward the concreteness of a natural object. This gap is a specific kind of "obscure zone of reality" in the transductive operation of individuation.

6.5. Differentiating Life from Machine

Transduction is the crucial concept in Simondon's critique of information theory and in his argument for distinguishing the living from the machine. He adapted the concept from the transduction in engineering and in biology, and we can find discussions on transduction in the transactions of the Macy Cybernetics Conferences. In this section, I will present the various meanings of transduction as explained by the conference participants, and how Simondon's conceptualization can be seen as a derivative of these meanings. Among these meanings is the biological notion of communication that elicits effects and reactions between the two ends, which grounds Simondon's critique of the notion of information in information theory. The elicitation of mutual effects lead to discovery or invention of biological organs in a process of self-organization that resolve problematics and tensions in organic growth. Simondon contends that an automaton, be it mechanical or informational, is not capable of such discovery or inventions.

The term "transduction" originates from science and technology and was brought up a number of times at the Macy conferences. During a discussion on memory and recalls, Norbert Wiener raised the idea of a mechanical recorder acting as a transducer capable of transcoding personal activities into recordable information (Pias, 1949/2003, p. 126). John Stroud followed up Wiener's explanation with an example of transduction:

An earphone is a device capable of transducing electromagnetic energies into sound waves (Pias, 1949/2003, p. 127).[124] In another meeting, Claude Shannon alludes to the transducer in his discussion of message translation between languages: “Physically we can think of a transducer which operates on the message to produce a translation of the message” (1950/2003, p. 271). In all these explanations, transduction refers to some form of mode conversion between activities, energies, and information.

In addition to mode conversion, Simondon’s concept of transduction also encompasses an operational notion of information, which is analogous to the biological communication as described by Herbert G. Birch and by W. Ross Ashby at other Cybernetics meetings. Birch defines biological communication “as the effect of the behavior of one organism upon the behavior of another organism” (Birch 1951, p.447). For example, “[i]f a starfish is placed in the environment of a scallop, it rather quickly elicits a flight reaction on the part of the scallop” (Birch, 1951/2003, p. 447).[125] While Birch did not identify this communication as “transduction,” Ashby employs the term to represent the biological communication in homeostasis. A homeostatic environment is “a transducer, as an operator that converts whatever action comes from the organism into some effect that goes back to the organism” (W. R. Ashby, 1952/2003a, p. 594). This notion of transduction as communication that elicits behavioral change is taken up by Simondon, for whom the operation of transduction “allows for signal of information to pass, but this passage, instead of being a simple conveyance of information, is integration or differentiation” (Simondon, 1964/2020, p. 171).

The distinction between the signal of information “as a simple conveyance of information” and as the seed that sparks activities of “integration or differentiation” is pivotal to Simondon’s critique of information theory. When “the living being evaluates its own action,” this evaluation cannot be reduced to a “simple consciousness of the discrepancy between the end and the result, and thus to a simple signal” (Simondon, 1964/2020, p. 172). A living being is one that undergoes transduction in both its interface to the external milieu and within its interiority where sub-individuals (e.g., cells within a

---

124 “Stroud: A typical transducer is an earphone. On one side you are putting in energies, pressure systems, and on the other side you are getting out some representation of it but in voltage. What comes out on the other side is supposed to represent what goes in but not necessarily in the same system of energy value, etc., but it has got to be a good representation” (Pias, 1949/2003, p. 127).

125 This example was discussed earlier in Section 3.2.

body) may incessantly merge and split: "The heterogeneous transductive characteristics only appear in the margins of this physical reality; on the contrary, interiority and exteriority are everywhere in the living being" (1964/2020, p. 172). In organic activities, the process of integrating sub-individuals into individuals (i.e., individuation) is followed by the process of differentiating the integrated unit from other individuals, and this loop would recurrently take place, inducing structural changes in individuals.

This argument identifies the limitation of Norbert Wiener's notion of negative feedbacks and of Ashby's design of the homeostat. A cybernetic automaton, such as Ashby's homeostat, cannot be a model for the living, for

> the automaton can only adapt in a manner convergent with a set of conditions by increasingly reducing the gap that exists between its action and its predetermined end; but it does not prevent [*sic*] and does not discover ends during its action, for it does not carry any veritable transduction since transduction is the expansion of an initially very restricted domain that increasingly takes on size and structure; biological species are endowed with this capacity of transduction due to which they can indefinitely expand. (Simondon, 1964/2020, p. 172)[126]

An automaton, built on cybernetic principles, cannot invent and discover its own goals, which have already been predetermined by its designers. Nor does it have the capacity to grow or reproduce organically, begetting an organization that has no continuity from the automaton's existing structure. Ashby's homeostat does not exhibit the "quantum nature of discontinuous action" (1964/2020, p. 172) that allows for the process of self-organization in organic growth. Life, in contrast, is "this mixture of continuous and discontinuous that is manifested in the regulative qualities which serve as a rapport between integration and differentiation" (1964/2020, p. 172). This opposition between the "quantum nature of discontinuous action" and "the continuous nature of the constructive knowledge of synthesis" (1964/2020, p. 172), this "mixture of the continuous and discontinuous" (1964/2020, p. 172), is the obscure, intermediate zone of activity. The "continuous" in the context of organic individuation seems to mean the safeguarding of stability and equilibrium that are essential for the survival of a living being.[127] The

[126] The word "prevent" is mistranslated in *Individuation in Light of Notions of Form and Information* (1964/2020, p. 172). It should be "invented."

[127] "From this point of view, it would be interesting to consider superior animal forms as arising from the *neotenization* of the inferior species in which the stage of individual life corresponds to the function of amplificative reproduction, whereas the stage of life *in colonies* corresponds to the continuous, homeostatic aspect. In superior species, individuals are the ones that live in society:

“discontinuous” is the “inventive” nature of growth in life, the transductive activities that do not logically proceed from the existing state and structure of the vital individual, but rather involves unpredictability and choices, whether that be the choices of nature or of a living subject. Life cannot be reduced to just the “continuous, homeostatic aspect” (Simondon, 1964/2020, p. 390). Instead, it is actually “conditioned by the recurrence of causality” between the continuous and the discontinuous (1964/2020, p. 173), between stability and inventive creativity.

6.6. Conclusion

In this chapter, I have argued that Simondon’s philosophy is both inspired by the ideas in cybernetics and, at the same time, a critique of cybernetics. As Yuk Hui explains, there are “two images” of cybernetics: “One is reductionist … The other is non-reductionist, in the sense of Simondon’s general allagmatic, which seeks genesis beyond any form of technological determinism” (Hui, 2019, sec. 44). Simondon developed the theories of individuation and concretization, as well as the concepts of recurrent causality and transduction, from the cybernetics ideas such as ontogenesis, feedback, and homeostasis. These theories and concepts embrace the themes of openness and complexity, which are shared by Wiener’s approach to scientific discovery to account for the irregularities that escape the closed system of classical analysis. At the same time, these theories and concepts are formulated to overcome the mechanistic thinking in cybernetics. Contrary to the cybernetic approach to treat ontogenesis as mechanical processes, individuation is both an operation of ontogenesis and an operation of pre-individuality, which makes creation and invention possible in the living. Cybernetic feedback, as a recurrence of information flow, is reconceptualized as the “recurrence of causality,” which indicates how two processes mutually and recurrently cause changes to each other. The term takes on the specific meaning of “an invention that *presupposes the problem to be resolved*” in technological concretization (Simondon, 1958/2016, p. 57 emphasis in original). Concretization is a theory about such inventive resolution of tensions and conflicts between technical milieux, which may be artificial or natural. It is analogous to the biological concept of homeostasis, as both are concerned with the resolution of tensions between milieux. But concretization is also an

---

the two stages and the two manners of being becoming simultaneous.” (Simondon, 1964/2020, p. 390) appropriation of ontogenesis, as both take on successive inventions and assemblages of structures and functions that solve, step by step, the internal problematic of an individual (1964/2020, p. 204). Such inventions of structures, whether in a living organism or in technical evolution, is neither deductive nor inductive, but transductive. Transduction is the crucial concept in distinguishing the living from the machine. Simondon adopts the term from its usage in sciences, which may convey the conversion between modes (activities, energies, or information) or a kind of biological communication. Unlike the communication in information theory, biological communication in living organisms elicits reactions and changes in one another. This concept of transduction becomes the basis for Simondon in how to differentiate life from machine. A machine such as Ashby's homeostat "does not carry any veritable transduction" whereas "biological species are endowed with this capacity of transduction" (1964/2020, p. 172). A machine built on cybernetic principles cannot invent and discover its own goals.

This critique of cybernetic machines seems particularly relevant to the question posed at the beginning of this chapter on whether AI can be creative. According to ChatGPT (see the beginning of this chapter), some people argue that "creativity can be defined more broadly as the ability to generate novel and useful outputs that are valued by humans. From this perspective, a machine such as DeepBach could be considered creative if it produces music that is novel, high-quality, and aesthetically pleasing to humans." In short, these people argue that AI can be considered creative because it can serve as a creative or aesthetic function for its users. But does this functional view of creativity necessarily imply that AI, empowered by deep learning, is capable of being creative and emotive? An argument built on Simondon's philosophy would refute any such a claim. As discussed in Chapter 5, AI empowered by deep learning can be characterized either as a pattern recognition engine with superhuman ability, or as an inductive algorithm based on the exhaustive search of complex hypotheses coupled with the elimination of false hypotheses with massive amount of data. Taking Simondon's philosophy as our basis, the question of AI's creativity can be rephrased as follows: Can the affordance of pattern recognition, or the training algorithm based on inductive logic, be reconceived as a transductive operation that can result in successive invention of new structures and functions to resolve sociotechnical tensions or problematics? While the examples of DeepFake or DeepBach, built using specific types of deep neural networks (e.g., Generative Adversarial Network (GAN), Recurrent Neural Network

(RNN), or Long Short-Term Memory (LSTM) Network), can successively generate new pictorial or musical structures, these structures are generated inductively, repeating patterns observed from the past with stochastic variations. These algorithms are not transductive, as they are not capable of inventing and creating new structures that convert contradictions into resolutions of the inner problematic in a technical or sociotechnical system. They can discover patterns and regenerate data embedded with such patterns, but they cannot create and invent like living organisms, which according to Simondon's philosophy undergoes transduction in both its interface to the external milieu and within its interiority.

Suppose we collect data on all kinds of creative acts and supervise machine learning to discover patterns of creativity. Could the model be trained to identify creative patterns like Simondon's transduction and concretization in technical inventions? Human creativity seems to involve subconscious acts capable of drawing relations from background information absorbed through fringe consciousness, and deep learning has been proven to be effective in replicating the function of fringe consciousness in applications such as language translation or autonomous driving. If deep learning AI can address Hubert Dreyfus' critique of fringe consciousness (see Section 1.4), what stops it from recognizing patterns of creativity? This argument is problematic on two fronts. First, transduction involves a creative introduction of a new perspective or a dimension of reality that is constructed from conflictual objects. Such an undiscovered dimension could not be uncovered in prior patterns. Second, it would be difficult to fathom the possibility of repetitive patterns that correlate this new dimension with the vast varieties of conflictual relations. For Simondon, it is the potentiality from pre-individuality between human imagination and the objects that make transduction, concretization, and recurrent causality possible. A machine learning model is a pattern-recognizing engine and does not share this pre-individuality. An AI enthusiast may then question, can a software program be written to exhaustively go through every object in the world and stops when it finds a combination in which components' conflicts are resolved like in concretization? First, this program would be unrelated to machine learning; it is a program written to realize Simondon's theories. Second, like the halting problem, the program does not know when to stop. It does not know when a certain combination is elegant enough that it should stop searching, even if it can recognize an elegant combination. Third, the search space is so large for every attempt to invent ($O(N^m)$[128] where N is the total number of beings in the universe and m is the number of components being put together) that it is practically impossible to produce any meaningful result.

The question of AI's emotivity, posed at the end of last chapter, can be addressed in a similar argument, as Simondon also applies his philosophy of individuation to understand human psyches. AI empowered by deep learning can certainly appear to behave emotionally, since such behavior can be implemented by generating the appropriate emotional reaction to many possible circumstances. It is capable of sentiment analysis, recognizing the feeling behind a facial or linguistic expression and regenerating the pattern of appropriate reactions to such expressions. But a robot or a computer simulation capable of imitating the emotional behavior of a human being ought not be characterized as an emotional being. In Simondon's philosophy, a person experiences emotion or affect because of the unresolved intensities from different levels of individuations, and affective intensities are resolved in the transductive operation of psychic individuation.[129] Such transductive operation is not commensurable with the deductive and inductive operation in machine-learning algorithms. A computer program like DeepBach can imitate the musical style of Johann Sebastian Bach in music generation, but it cannot compose new forms of music in response to a changing political and social environment. It is not capable of inventing like Arnold Schoenberg's twelve-tones techniques, or composing music that react to political suppression like the anti-Stalin messages in Dmitri Shostakovich's symphonies[130]. Therefore, however impressive AI applications such as DeepBach or ChatGPT may appear to their users, they are no more than engines with the technical affordances of pattern recognition and of stochastic regeneration of discovered patterns.

Even in the argument that "DeepBach could be considered creative if it produces music that is novel, high-quality, and aesthetically pleasing to humans," the function of regenerating musical patterns may appear creative and emotive only through the

---

[128] $O(N)$ is the notation commonly used in computer science to denote the "order of N." If an algorithm is in the order of N^m, which can be written as $O(N^m)$, it means that the algorithm takes exponentially longer time to process.

[129] I elaborated this point in my paper "The Dilemma of Openness in Social Robot" (2019). See also David Scott's *Gilbert Simondon's Psychic and Collective Individuation* (2014).

[130] This is the view adopted by the documentary film *The War Symphonies: Shostakovich Against Stalin* (Weinstein, 1997).

appreciation of a human listener. In general, technical inventions or sociotechnical innovations can acquire the meaning of creativity or emotionality only with the human as its counterpart. In Simondon's philosophy, the potentiality of creativeness and the intensities of affect originate from the pre-individual reality shared between the human and its sociotechnical milieu. This concept of pre-individuality, which basically means the reservoir of potentials that will not be exhausted (Hui, 2019, sec. 35), is characterized by some as speculative ontology (Seibt & Rodogno, 2019). But this seemingly speculative concept has its physical basis from Simondon's understanding of quantum physics and crystallization. As I will explain in the next chapter, the discontinuities in the atomic and quantum realm become the physical basis for Simondon's paradigm of discontinuous figures and continuous ground, and this figure-and-ground paradigm is significant to his conceptualization of recurrent causality, transduction and concretization.

Chapter 7.

The Potentiality of Pre-Individual Reality

Me: Can you explain how Simondon's concept of pre-individuality is tied to quantum theory?

ChatGPT: Simondon's concept of pre-individuality is tied to quantum theory in several ways. First, Simondon's concept of pre-individuality is rooted in the idea that entities in the world, including technological objects, are not pre-existing but instead emerge through a process of individuation. Similarly, in quantum theory, particles and systems are not seen as pre-existing entities that interact with each other but instead are seen as emergent entities that come into being through a process of interaction with their environment.

Simondon's concept of pre-individuality is also tied to quantum theory through the idea of entanglement. In quantum theory, particles can become entangled, meaning that the properties of one particle become dependent on the properties of another particle, regardless of their distance from each other. Similarly, in Simondon's philosophy, pre-individual entities are not yet fully differentiated, meaning that they are dependent on and entangled with other entities in their environment.

Furthermore, Simondon's concept of pre-individuality is tied to quantum theory through the idea of superposition. In quantum theory, particles can exist in a state of superposition, meaning that they exist in multiple states simultaneously. Similarly, in Simondon's philosophy, pre-individual entities exist in a state of potentiality, meaning that they have the capacity to become differentiated and actualized in multiple ways.

Overall, Simondon's concept of pre-individuality is tied to quantum theory through the ideas of emergence, entanglement, and superposition. Both concepts suggest that entities in the world are not pre-existing but instead emerge through a process of interaction with their environment and that they exist in a state of potentiality or superposition before becoming fully differentiated.

7.1. Introduction

In the last chapter, I described in what ways Simondon's philosophy was inspired by cybernetics and explained how it may overcome the problem of mechanizing life in cybernetics. The differentiation between Simondon's philosophy and cybernetics served as our basis for differentiating computational creativity from the creativity of the living, which Simondon theorizes in the notions of concretization for technical inventions and of transduction for physical, biological, psychic, and collective individuation. But critics may question, are all these philosophical concepts simply the product of his own speculations? Indeed, there are people who would characterize Simondon's theories as speculative ontology (e.g., Seibt & Rodogno, 2019). But as I will contend in this chapter, Simondon's theories actually find their basis in the scientific theories of physical phenomena, which have become radically different since the discoveries and theoretical breakthroughs during the 19th and 20th century. Whereas Aristotle's model of potentiality was life (Feenberg, 2023, p. 85), Simondon's model of potentiality was based on modern physical sciences. Quantum theories explain observations obtained via more advanced scientific instruments that can capture behavior of waves and subatomic particles, while solid-state physics describes the 3-dimensional propagation of a molecular lattice in crystallization. These twentieth-century scientific theories attempt to explain observations that violate classical logic and the common sense of everyday life. Rather than disregarding scientific breakthroughs as epistemologically problematic,[131] Simondon sees the necessity of a new philosophy that is consistent with these seemingly non-sensical theories about the physical world.

This is why, as Andrew Iliadias has suggested, it is imperative to interpret Simondon by understanding the relations between his philosophical concepts and the various theories in modern sciences:

> What Deleuze did not point out, and what many English readers of Simondon have heretofore failed to pick up on, is that in articulating this new philosophy Simondon was simultaneously engaged in conversation with some of the most technically advanced scientists, engineers, and mathematicians of the twentieth century. Any real understanding of Simondon's approach to individuation – most central of all Simondonian concepts – must acknowledge the privileged position that Simondon gave

[131] I am referring to the strong programme in the sociology of scientific knowledge.

> to notions from within engineering, physics, and especially cybernetics in his original ontology. (2013, p. 1)

I have explored the significance of cybernetics and engineering (such as a concise understanding of the Guimbal engine) in the last chapter. In this chapter, I will look into the extensive passages on quantum physics and crystallization in *ILNFI*. Doing so will substantiate my contention that Simondon's philosophy is not any more speculative than Aristotle. Simondon simply treats the peculiar and mysterious subatomic behavior as seriously as the way Aristotle treats ordinary experience and human intuition applied to large objects.

These passages on advanced physics, however, seem to assume that readers are knowledgeable in quantum theories and solid-state physics. Readers with no background knowledge of these theories would find them incomprehensible. By not taking these passages seriously, readers would unknowingly turn Simondon's modeling of physical phenomena into a fascinating but unfounded speculative philosophy. Yet, hope is not all lost for people who have not studied particle physics in university. It is possible to develop an intuitive understanding of the uncanny subatomic behavior and of the molecular/atomic activities in crystallization without going into complex mathematics equations.[132] To do so, I will draw from the explanations in *The Feynman Lectures on Physics* (Feynman et al., 2011), which is a physics textbook based on lectures given by Richard Feynman to undergraduate students from 1961 to 1963. Feynman is a Nobel laureate who has sometimes been called "The Great Explainer" (LeVine, 2010), and is well-known for his knack of giving intuitive explanation behind the complex mathematical theories in physics. These lectures came after Simondon's defence of his doctoral dissertation in 1958, which was later turned into the publications of *MEOT* and *ILNFI*. So clearly, his knowledge of physics in *ILNFI* did not come from Feynman's lectures. Nevertheless, an intuitive grasp of quantum theory and of crystalline lattices through these lectures can enlighten our interpretation of Simondon's theory of individuation and his notion of pre-individual reality.

This enlightened interpretation would reveal the significance of the figure-and-ground paradigm for Simondon's critique of technological alienation and why

---

[132] I am also not interested in the latest advances in these fields, as the goal is to simply understand Simondon.

concretization is the means to overcome such alienation. Simondon's theory of physical individuation and pre-individual reality are attempts to make sense of the quantum behavior of subatomic particles and of the crystalline activities in the molecular and atomic realm.[133] In *ILNFI*, Simondon talks about quantum mechanics and wave mechanics as two ways of expressing pre-individuality:

> Perhaps it would be in this sense that we could see the convergence of these two new theories, that of quanta and that of wave mechanics …; they could be envisioned as two ways of expressing the pre-individual through the different manifestations in which it intervenes as pre-individual. Below the continuous and the discontinuous, there is the quantic and the metastable complementary (the more than unity), which is the true pre-individual" (Simondon, 1964/2020, p. 6).

Acquiring such intuitions is essential to the understanding of the pre-individual reality as the being "below" the continuous and the discontinuous in the observable behavior of electrons and of crystallization, and how the "charge of pre-individual reality" are transduced to bring about the formation, structuration, and organization of individuals. Simondon portrays the discontinuous and the continuous in a figure-and-ground paradigm, the discontinuous being the figure and the continuous being the ground. He then proposes that relations between the discontinuous and the continuous, at all levels of individuation, subsist in a recurrence of causality between the figure and the ground. As I will explain in this chapter, technological alienation for Simondon is a consequence of the blockage between the figure and the ground that would inhibit the operation of individuation, and concretization is a scheme that can possibly overcome technological alienation.

In the following, I will begin by presenting an intuitive explanation of quantum theory and crystallization based on *The Feynman Lectures on Physics*. I will then describe the significance of this understanding for Simondon's conceptualization of pre-individuality. This can help us see why recurrent causality is about the relation between the discontinuous and the continuous, and consequentially, between the figure and the ground. I will further elaborate on Simondon's idea that technological alienation comes

[133] His elaboration on vital and biological individuation are just as important, but it is much easier to understand for non-biologist. In *ILNFI*, he brings up biological entities such as sea anemone to problematize the notion of individual due to the unclear boundaries and couplings of many biological entities. Because of its comprehensibility, I will not explain Simondon's theory on vital individuation in this dissertation.

from the blockage between the figure and the ground, and on how concretization is the means for overcoming this alienation.

7.2. Quantum Mechanics

Simondon's philosophy is difficult to understand due to the vagueness of his abstract concepts that can be too open to interpretations. Understanding his detailed scientific and technical illustrations can clarify his intended meanings, but readers not well-trained in advanced physics cannot gain much from his writing on quantum theory and wave-particle duality. Yet, the uncanny behavior of subatomic activities, counterintuitive and contrary to our everyday experience of large objects, is also what necessitates a philosophy distinct from the classical system of logic for Simondon. This classical system was formulated by ancient thinkers who only had direct experiences with the physics of large objects. Simondon attempted to formulate a philosophy distinct from classical philosophy by taking new scientific discoveries on subatomic activities into account.

As mentioned earlier, in our endeavor to understand Simondon's passage on quantum mechanics, we will cross-read Simondon with *The Feynman Lectures on Physics* (Feynman et al., 2011). As Feynman explains,

> Because atomic behavior is so unlike ordinary experience, it is very difficult to get used to, and it appears peculiar and mysterious to everyone—both to the novice and to the experienced physicist. Even the experts do not understand it the way they would like to, and it is perfectly reasonable that they should not, because all of direct, human experience and of human intuition applies to large objects. We know how large objects will act, but things on a small scale just do not act that way. So we have to learn about them in a sort of abstract or imaginative fashion and not by connection with our direct experience. (Feynman et al., 2011, bk. III, Chapter 1-1)

Feynman then illustrates the peculiar and mysterious subatomic behavior in a "thought experiment,"[134] showing how electrons may behave like particles, like waves, like both, or like neither. In this thought experiment, there is an electron gun and a movable electron detector, which can move up and down along the backstop (see Figure 20). The

[134] A "thought experiment" is a hypothetical experiment in which "we know the results that would be obtained because there are many experiments that have been done" (Feynman et al., 2011, bk. III, Chapter 1-4).

detector, which can be a Geiger counter connected to a loudspeaker, would produce a "click" sound upon each arrival of an electron. With the detector at different positions, the rate at which the clicks appear is faster or slower, but the size (loudness) of each click is always the same. There are no "half-clicks." And if two separate detectors are placed at the backstop, only one of them would click but never both at the same time. We can therefore conclude that "[e]lectrons always arrive in identical lumps" and behave like particles (Feynman et al., 2011, bk. III, Chaper 1-2), and would expect the behavior of an electron gun emitting electrons to be just like that of a machine gun shooting a stream of bullets (See Figure 18).

Now, as we can see in Figure 20, there are two holes in the wall that allow electrons to go through, and we can observe the probability distribution of where the electrons may land with either hole open or both holes open. What is peculiar here is that this probability distribution is similar to the probability distribution in a set-up with a "wave source," such as a speaker generating sound waves (see Figure 19), and different from the probability distribution if the set-up has a "particle source," such as guns shooting bullets (see Figure 18). With guns shooting bullets, the probability distribution with both holes opened is equal to the sum of the probability distributions with either hole opened (see P_{12} in Figure 18). With a wave source, this is not the case as there are interferences when the waves through the two holes arrive at the detectors "out of phase" (e.g., with a phase difference of pi), and the waves would "interfere destructively" (see I_{12} in Figure 19). With an electron gun as the source, the probability distribution also exhibits wave-like interference when both holes are opened (see P'_{12} in Figure 20), even though each electron appears to be an individual corpuscle like bullets, based on the unity of each "click" sound produced via the Geiger counter. Feynman thus gives the following conclusion: "The electrons arrive in lumps, like particles, and the probability of arrival of these lumps is distributed like the distribution of intensity of a wave. It is in this sense that an electron behaves 'sometimes like a particle and sometimes like a wave'" (Feynman et al., 2011, bk. III, Chapter 1-5).

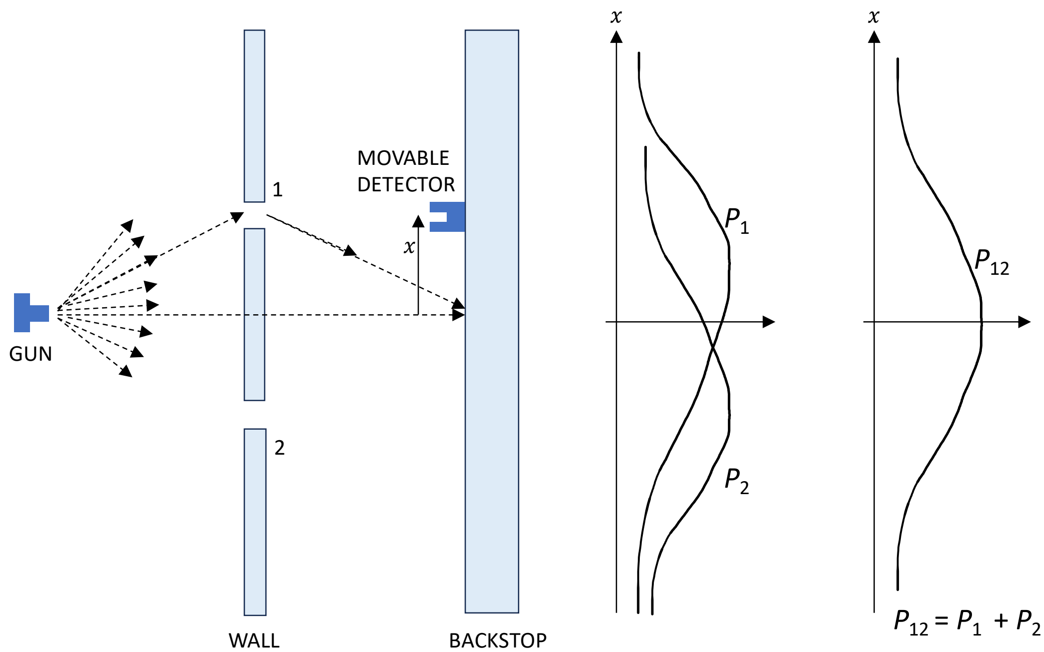


Figure 18: Interference experiment with bullets[135]

To further determine whether electrons behave more like particles or more like waves, Feynman examines the proposition that “each electron *either* goes through hole 1 *or* it goes through hole 2” (Feynman et al., 2011, bk. III, Chapter 1.5). If this proposition is true, then the number of electrons that arrive at a particular point on the backstop should be equal to the number arriving via hole 1 plus the number arriving via hole 2, and the probability distribution with both holes opened should not exhibit wave-like interference. But since we already know about the wave-like interference, then the proposition must be false. To confirm this logical deduction experimentally, we can place a very strong light source behind the wall and between the two holes, as shown in Figure 21, such that when an electron passes by, it would scatter some light to our eyes. So if an electron goes through hole 1, then a flash of light would come in the vicinity near hole 1. If it goes through hole 2, then a flash of light would come in the vicinity near hole 2. Running the experiment with this light-source set-up, we would see that “every time that we hear a ‘click’ from our electron detector (at the backstop), we also see a flash of light either near hole 1 or near hole 2, but never both at once” (Feynman et al., 2011, bk. III, Chapter 1.4)! To our surprise, contrary to our deductive reasoning, the proposition

[135] Figure 18 to Figure 21 are adopted from *The Feynman Lectures on Physics* (Feynman et al., 2011, bk. III, Chapter 1).

that electrons go through either hole 1 or hole 2 is actually true according to our experiment. Now, in order to find out what is wrong with this deductive reasoning, we can measure the probability distribution as before, counting the number of electron arrivals at each position on the backstop, except that this time we also keep track of which hole each electron has gone through. So we can get a probability distribution for electrons going through hole 1 (P_1' in Figure 21), which would look the same as the probability distribution when hole 2 is blocked off as in Figure 20 (P_1). And we can get a probability distribution for electrons going through hole 2 (P_2' in Figure 21), which would look the same as the probability distribution when hole 1 is blocked off (P_2). We can also get the total probability distribution (P_{12}' in Figure 21) by counting the total number of electrons arriving at each position. And strangely enough, the total probability distribution is now equal the sum of the probability distribution for electrons going through hole 1 and that for electrons going through hole 2 ($P_{12}' = P_1' + P_2'$). In this experimental set-up with the light source, there is actually no sign of interference!

Thus it appears that the act of looking at electrons, via the light source, would change their behavior, such that they would behave completely like particles. Could it be the light source that disturb their behavior? What happens if we adjust the light source, making it dimmer or adjusting the wavelength? Would it reduce the disturbance by the light source on the electrons? We can also try this out experimentally, and it turns out that, by dimming the light source, some electrons can be seen while others may pass by the hole unseen (that is when “clicks” are heard with no “flash of light” near either holes). Only those electrons that we can see would fall onto a particle-like probability distribution, whereas those that escape our eyes would fall onto a wave-like probability distribution. Therefore, “it is impossible to arrange the light in such a way that one can tell which hole the electron went through, and at the same time not disturb the pattern” (Feynman et al., 2011, Chapter III 1-6). This is Heisenberg’s uncertainty principle, and “[n]o one has ever found (or even thought of) a way around the uncertainty principle. So we must assume that it describes a basic characteristic of nature” (Feynman et al., 2011, Chapter III 1-6).

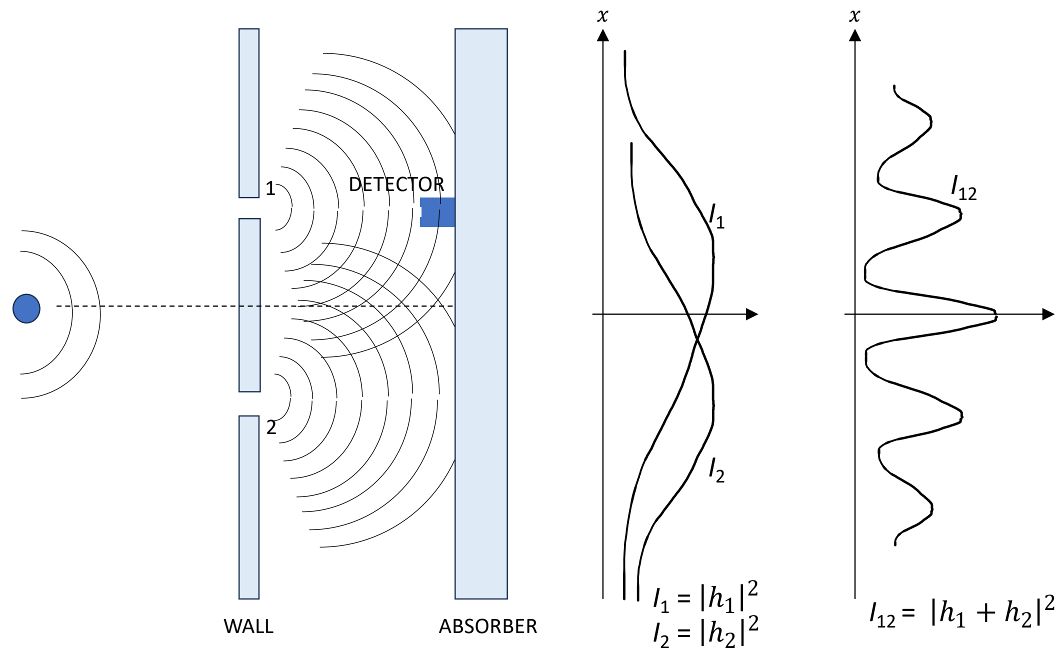


Figure 19: Interference experiment with water waves

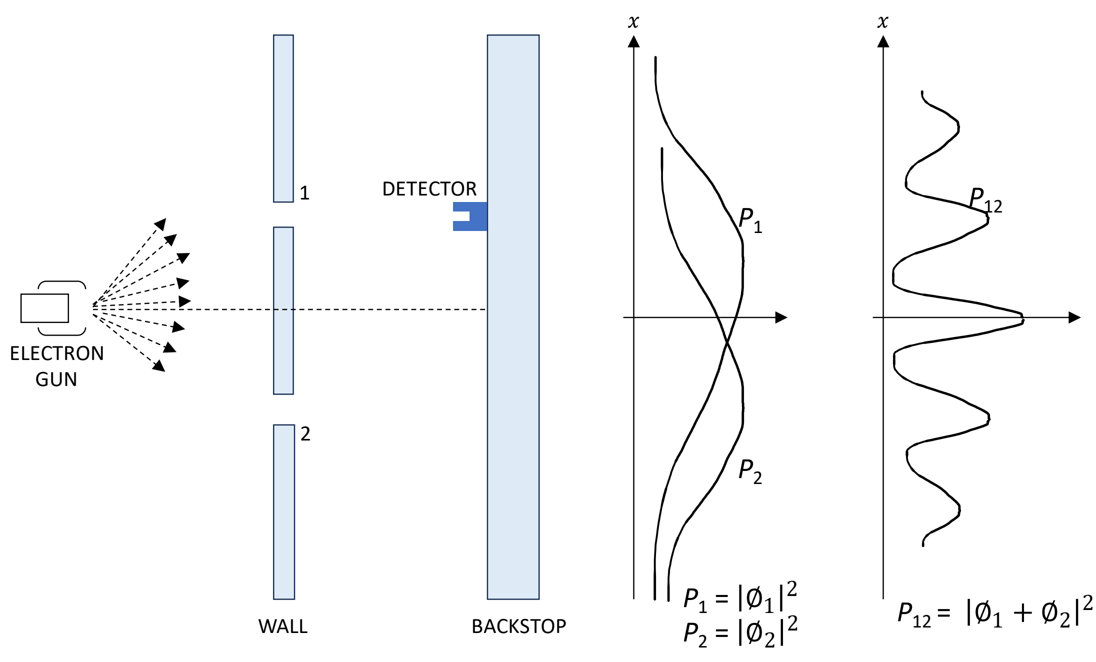


Figure 20: Interference experiment with electrons

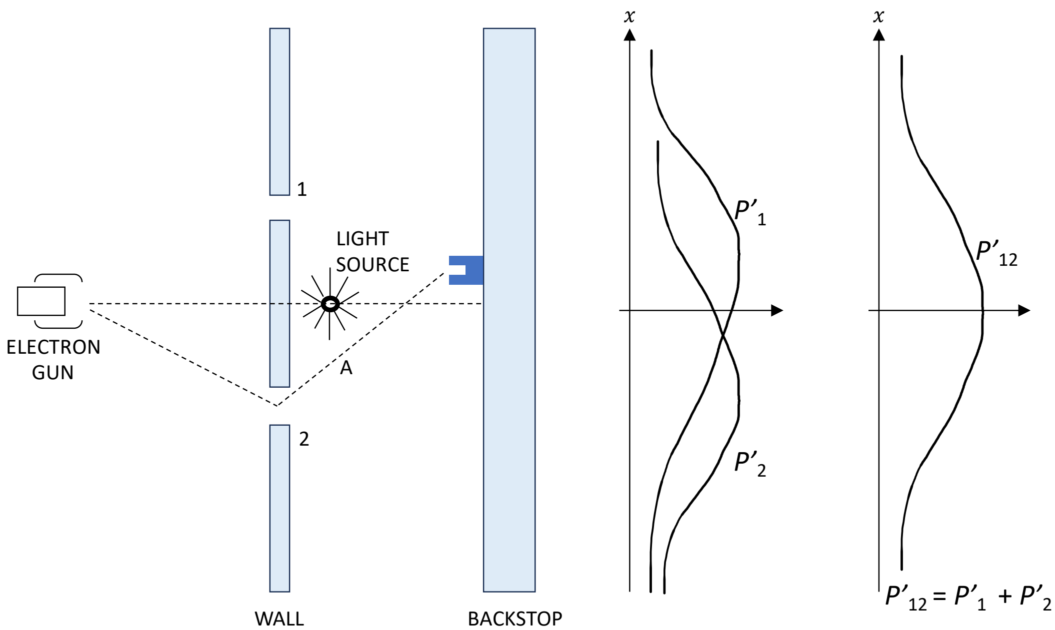


Figure 21: A different electron experiment

To sum it up, electrons behave like individual particles when they are observed via the Geiger detector, which converts or "transduces" its lump of energy into audible sound wave. These same electrons exhibit wave-like behavior when the "clicks" are counted to generate a probability density curve representing where they may land on the backstop. So they exhibit both particle-like behavior and wave-like behavior. But when a light source is added, allowing human eyes to trace the route that each electron passes through, from the electron gun through a hole to the backstop, the probability distribution of where electrons land is like that of large particles.

This type of uncanny behavior in the subatomic world, which is peculiar and mysterious because our intuitions are based on direct experience with the physics of large objects, became for Simondon a "new path for grasping the reality of the individual" (1964/2020, p. 136). This new path "opens up with quantum theory, whose power of transductivity is so great that it allows for the establishment of a viable relation between an inductive physics of the discontinuous and a deductive energetic theory of the continuous" (1964/2020, p. 136). The word "transductivity" is appropriated from science and technology,[136] and the term was also frequently employed by participants of the

[136] See footnote 124.

Cybernetics Conference.[137] Simondon appropriates this scientific and technical term to his philosophy to account for phenomena that cannot be explained by direct causal relation. Such phenomena signify the reversal of the law of entropy, giving rise to the formation of individuals or organized structures without any direct, visible causes. In Simondon's philosophy, negentropic phenomena[138] are the result of transduction, an operation where the pre-individual charge, from a mode of reality prior to the becoming of the individuals (hence called pre-individual reality), is transduced to bring about the formation of individuals or organized structures.

We can use Simondon's concepts of transduction and pre-individual reality to explain the uncanny behavior of the electrons in Feynman's "thought" experiment. When electrons are observed with the Geiger counter, we can say that they are transduced from the form of waves to the form of particles. When electrons are observed with the light source, the disturbance caused by the energy field of the light source transduces electrons from the form of waves to the form of particles. It appears that, when electrons are observed by whatever means, they are transduced from the continuous form of waves into the discontinuous form of individual particles. From this, Simondon concludes that between the discontinuous and continuous forms of reality exists some obscure, intermediate zone of reality where transductive activities operate. The physics of particles and of waves in large objects fails to explain the subatomic world because particles and waves are only "extreme terms" that overlook this intermediate zone of reality. This intermediate zone is analogous to the one in the process of brickmaking, in which form and matter are only extreme terms that do not account for the intermediate zone of chemical reactions in the formed matter of clay and the material form of the mould.[139]

---

[137] See Sections 4.2 and 6.5.

[138] The term "negentropy" is coined by Bernard Stiegler. Negentropy reverses the second law of thermodynamics, which states that "the total entropy of a system either increases or remains constant in any spontaneous process; it never decreases" (Urone et al., 2020, Chapters 12–3). Note that Stiegler often uses this term "negentropy" while Simondon never uses it.

[139] Pascal Chabot gives a succinct summary of how Simondon critiques hylemorphism with the example of brickmaking in *The Philosophy of Simondon* (2013, pp. 75–78)

For Simondon, the uncanny behavior of particle-wave duality of electrons can be explained with the theory of transduction in an intermediate zone of reality. This understanding helps us interpret the meaning of the following passage in *ILNFI*:

> In the end, we could ask ourselves whether or not, instead of being capable of entering into the framework of an indeterministic physics or that of deterministic physics, we should consider the theory of singularities as the foundation for a new representation of the real that encompasses these two as particular cases and that should be called the theory of transductive time or the theory of the phases of being. This definition of a new manner of thinking becoming, which calls for determinism and indeterminism as borderline cases, applies to other domains of reality than that of elementary corpuscle … (1964/2020, p. 154)

The framework of a "deterministic physics" likely denotes the Newtonian mechanics that can predict the behavior of particles, while the framework of an "indeterministic physics" likely denotes Maxwell's waves theory. For Simondon, both are "borderline cases" that do not apply to the domain of "elementary corpuscle," a domain of reality that can be accounted for by "a new representation of the real that encompasses these two as particular cases," and this new representation "should be called the theory of transductive time or the theory of the phases of being."

7.3. Crystallization

As with his writing on quantum theory, Simondon wants to bring into relation the discontinuous and the continuous in the process of crystallization. Many commentaries of Simondon present crystallization as a prototypical example of his philosophy of individuation, but their accounts are lacking in explaining the discontinuities that Simondon emphasizes in the process of crystallization.[140] Some details in the scientific knowledge about crystals are significant for understanding Simondon but are easy to overlook. In this section, I will describe the scientific details and then examine the parallel between the continuous, the discontinuous, and the intermediate zone of reality in crystallization and the preceding account of transduction in subatomic activities.

Crystal is a type of solid material with a repetitive pattern of atoms or molecules in a 3-dimensional lattice. This pattern is formed when atoms do not move around very

[140] For instance, neither Chabot (2013) nor Combes (2013) elaborate much about what Simondon means by discontinuities in crystallization.

much and arrange themselves in a configuration with as low an energy as possible (Feynman et al., 2011, bk. II, Chapter 30-1). So the process of crystallization involves the realignment of atoms or molecules, binding themselves to a crystalline lattice and propagating its 3-dimensional pattern. A seed crystal is simply a small piece of a crystal or polycrystal material of the same material, constituted of a miniature base pattern that can be expanded, like a wallpaper (See Figure 22). The physical properties of a crystal vary depending on the kind of bonds between the atoms or molecules that make up the lattice. In a diamond, the carbon atoms have covalent bonds in all four directions to the nearest neighbors, and the crystal is very hard to break. A sugar crystal and paraffin are molecular crystals with weak attractions between molecules. So sugar crystals are easy to break. In metals, the bonding is of an entirely different kind, as the bonding is not between adjacent atoms. Each atom contributes an electron to a universal pool of electrons, and the positive ions reside in the sea of negative electrons, which holds the ions together like some kind of glue (Feynman et al., 2011, Chapters 30–1).

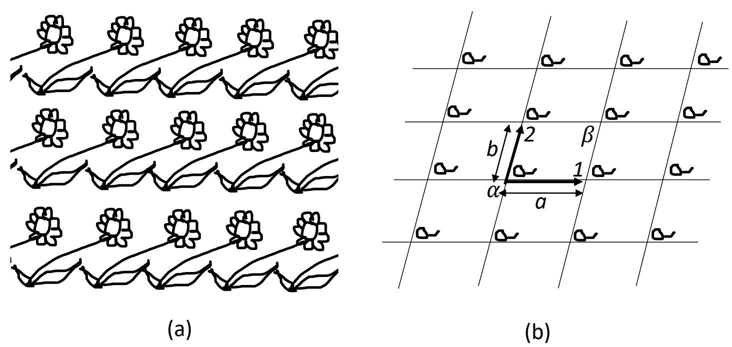


Figure 22: A repeating pattern in two dimensions[141]

A couple of characteristics of crystalline lattice are particularly relevant to our understanding of Simondon. First, there is a physical discontinuity between the molecules in the lattice. They are not touching one another, and unlike chemicals that are held more strongly by ionic bonds or covalent bonds, crystals can be held together by "nonbonded interactions" (Dlott, 2003, Chapter 3.2). According to Simondon, the

> individuation that we will characterize through the example of the crystal cannot exist without an elementary discontinuity on a more restricted scale; it takes an edifice of atoms to constitute a crystalline lattice, and this

[141] The figure is adopted from *The Feynman Lectures on Physics* (Feynman et al., 2011, bk. II, Chapter 30-1).

> structuration would be very difficult to conceive without an elementary discontinuity. (1964/2020, p. 90)

The "elementary discontinuity on a more restricted scale" likely refers to the "non-bonded" interactions through which crystalline molecules or atoms may be held together. Second, crystallization is a process that reorganizes an amorphous milieu of molecules (the metastable solution). The process is therefore negentropic. The introduction of a crystalline germ, which is basically a lattice of molecules in a smaller scale, supplies the pattern for the process of crystallization. In this process, specific molecules in a supersaturated crystalline solution will bind themselves to the lattice, propagating the repeated pattern of the germ.

This scientific knowledge of crystals as lattices and their propagation of repeating patterns is helpful for understanding Simondon's writing on the individuation of crystals:

> the structural germ must bring with it a structure corresponding to the crystalline system in which the amorphous substance can crystallize; the crystalline germ does not have the same chemical nature as the amorphous crystallizable substance, but there must be an identity between the two crystalline systems in order for the apprehension of the potential energy contained in the amorphous substance to be carried out. … The individuation of a system essentially results from the meeting of a mainly structural condition and a mainly energetic condition (1964/2020, p. 80).

Transduction results from the encounter of a structural condition (the repeated patterns in lattices) with an energetic condition (supersaturated crystalline solution[142]). This encounter brings about the reorganization of molecules in the growth of a crystalline lattice. The potential energy for such a reorganization has always been latent in the energetic condition in the "amorphous crystallizable substance," but it would remain latent until its encounter with a structural germ that shares "an identity" with it. This shared identity presumably implies that the chemical constituents of the crystalline seeds is the solute that exceeds the concentration specified by the solubility in a supersaturated solution. The encounter between the two extreme conditions transduces the latent energy in the supersaturated solution to bring about the reorganization of ions or molecules that bind themselves onto the crystalline lattice.

[142] A solution is supersaturated when the concentration of a solute exceeds the concentration specified by the solubility. This metastable state may be brought back to equilibrium by forcing the excess of solute to separate from the solution ("Supersaturation," 2022).

Simondon further explains, “[t]his supposes that individuation exists on an intermediate level between the order of magnitude of the particulate elements and that of the molar ensemble of the complete system” (1964/2020, p. 94). Accordingly, there are two orders of magnitude, or two disparate domains of reality. One domain of reality is the order of the particulate elements. The other domain of reality is the order of the solid crystal such as diamonds, sugar, or metals. Individuation is an operation at the intermediate level between the two orders of magnitude, where the discontinuous interacts with the continuous: “[I]ndividuation is initiated on the level at which the discontinuous of the singular molecule is capable … of modulating an energy whose support is already a part of the continuum in the population of randomly arranged molecules” (1964/2020, p. 94). Transduction is then an operation at this intermediate level where an energy of the continuum is modulated by singular molecules to form the “molar ensemble.” In this process, “the tension between incompatible—as yet unrelated—dimensions of being” is transduced into another form of energy, one that “[integrates] disparity and difference into a coordinated system” (“A Short List of Gilbert Simondon’s Vocabulary,” 2007).

7.4. Recurrent Causality – Figure and Ground

In both the uncanny behavior of electrons and in the edifice of molecules or atoms in crystalline lattices, Simondon identifies an intermediate level of activities between the discontinuous (subatomic particles or atoms and molecules constituting a crystal) and the continuous (waves or supersaturated solution). He models this relation between the discontinuous and the continuous at all levels of individuation as a recurrence of causality between the figure (discontinuous) and the ground (continuous). In vital individuation, the appearance of differentiated organic individuals corresponds to the appearance of discontinuous figures, and the continuum of physical and living milieu from which emerge the figures corresponds to the metastable ground. Simondon then extends this concept of recurrent causality to his analysis of technology.

In this figure-and-ground paradigm, the technical objects are the figures, whereas the natural, psychic, or the social milieu serve as the ground of metastable continuum. “[Psychic ground] is the milieu associated with a systematics of forms that institutes relations of recurrent causality between these forms and that which causes recastings of the system of forms taken as an ensemble” (1958/2016, p. 61). When it is free from alienation, the psychic ground of affective intensities never stops channelling the energy that would cause the "recastings of the system of forms," which belong to the psychic ground themselves. When the channels of this recurrence of causality between structural forms of the figure and the energetic potentials of the ground are blocked or "short-circuited,"[143] the ground becomes alienated from the figures: "Alienation is the break between ground and forms in psychic life: the associated milieu no longer regulates the dynamism of forms" (1958/2016, p. 61). In the realm of technology, forms in psychic life can be associated with the technical schemas, which then subsist physically as technical objects or mentally as schemas within human psyches. In technological alienation, the dynamisms of technical schemas are no longer regulated by the psychic or natural milieu. Therefore, technological alienation can be subverted if the broken channel between ground and forms can be re-established, allowing the associated milieu of psychic ground or the ground of nature to "[regulate] the dynamism of forms," which is the same as the dynamism of technical schemas.

Understanding Simondon's recurrence of causality as a feedback loop between figure and ground further reveals the essence of Simondon's critique of Wiener's cybernetics. In this critique, cybernetics advocates a theory of automata built on feedback loops while excluding the ground of pre-individual charge from the milieu. Automatons cannot individuate and are lifeless in themselves, because their goal-oriented recursivity, a negative feedback loop that aims to reduce the distance between the current state and its final target, is based on pure mathematical and algorithmic operations that preclude any relations with a pre-individual ground of random, amorphous continuum filled with tensions, randomness, and disparations. When human activities are primarily conducted in systems made up of automatons, participating humans would be alienated from the ground of natural, psychic, or social milieu.[144]

---

[143] The term "short-circuited" is used by Antoinette Rouvroy and Thomas Berns (2013) as well as by Stiegler (2016) in their Simondonian critique of algorithmic governance and a society proliferated with automations.

[144] Yuk Hui also identifies with this crux of the problem with modern technology as one about the figure of technology detached from the ground, or becoming the ground itself: "As Schelling attempted to show, evil emerges when the figure is taking over the ground (again like Figure-Ground in Gestalt psychology), when the self-will takes over the universal will; seeking a solution in the self-will is an affirmation of the perversion of the ground, the perpetual loss of the universal will" (Hui, 2019, sec. 7). Hui's attempt to combat this "perversion of the ground" led to his conceptualization of cosmotechnics, which draws on non-western technics from different indigenous cultures in order to escape from the enframing of the calculable: "However, when

Even though our contemporary digital milieu is becoming increasingly automated, to the extent that it is seemingly taking over the status of the ground in people's lives[145], the ground with pre-individual charge will always subsist as long as there remains a world of nature, of human lives, and of social lives. If Simondon is correct that the source of alienation is the broken channel between the figure and the ground filled with pre-individual potentials, our approach to oppose alienation can come in the form of re-establishing the rapport between technics and the metastable milieu of physical nature, human psyches, and the collectives of transindividual relations.

7.5. Conclusion

In this chapter, I have drawn from the lecture series on physics by Richard Feynman, "the Great explainer," to present non-mathematical explanations on the behavior of subatomic activities in quantum mechanics and the molecular and atomic behavior in the formation of crystalline lattices. Feynman's thought experiment illustrates how electrons can behave sometimes like particles and sometimes like waves, depending on the experimental set-up. If there is no attempt in the set-up to observe which of the two holes the electrons go through, the electrons would behave like wave. If there are attempts to observe which of the two holes they go through, via the Geiger counter or the light source, the electrons would behave like particles. In Simondon's philosophy, the Geiger counter or the light source transduces electrons from the form of waves to the form of particles, from a continuous form to a discontinuous form. These transductive activities operate in an obscure, intermediate zone of pre-individual reality, in which particles and waves are the extreme forms.

I then explain the process of crystallization at a molecular level, which Simondon also theorizes as transductive activities in an intermediate zone of pre-individual reality. Crystallization is a process that reorganizes an amorphous milieu of molecules in a

---

technology detaches itself from this balance of figure/ground and becomes its own ground, as well as the ground of other domains, we will have to resituate it in a new episteme and transform it from within according to different epistemologies. This is also the reason for which we must search for the ground of technology. This was also my motivation in developing the concept of cosmotechnics as an attempt to open up the question of technology: We don't have only one technology (as figure) and one cosmology (as ground), but rather multiple cosmotechnics containing different dynamics between the moral and the cosmos" (Hui, 2019, sec. 38).

[145] See the footnote 144 on how Hui identifies the problem of computational technology as the perversion of the figure taking over the ground.

metastable solution. A crystalline germ is a small lattice of atoms or molecules in a particular configuration, and when it is introduced to an amorphous milieu of molecules, its configuration will serve as the pattern for the growth of the lattice. Randomly arranged molecules will be re-organized around the small lattice, propagating its growth outward by repeating the pattern of the germ's molecular configuration. The structure of molecules will consequentially become a large crystalline lattice. In *ILNFI*, Simondon describes this operation as the discontinuous molecules transductively modulating the energy from the continuum of randomly arranged molecules.

The paradigm of transducing the continuous into the formation of the discontinuous turns out to be fundamental to Simondon's concept of recurrent causality. He models the transductive relations between the continuous and the discontinuous as a recurrence of causality between the figure and the ground. This figure-and-ground paradigm is applicable to his thoughts on vital (biological) individuation as well as to his analysis of technology. Technical schemas are the figures, whereas the natural, psychic, or the social milieu is the ground of metastable continuum. Freed from alienation, the psychic ground of affective intensities would channel the energy for the transductive invention of new technology. Simondon theorizes technological invention in analogy to the transduction of electrons from waves into particles, or in analogy to the transduction of crystalline solution that brings about the appearance of crystals. Technological alienation is the result of a break between the figure of technical schemas and the ground of the natural, psychic, or the social milieu. Re-establishing the broken channel can therefore subvert technological alienation.

As mentioned in Chapter 6, the concept of recurrent causality is both an appropriation and a critique of cybernetics feedbacks. They share the aim of exploring the openness and the complexity in disorderly systems across heterogeneous environments. At the same time, cybernetics feedbacks operate purely as some technical schemas, which are designed to integrate the contingency in nature and in humans as technical modules that can be controlled.[146] In this framework of cybernetics feedback, the figure of technical schemas remains detached from the ground of nature and humans. Because deep learning AI operates within this framework, it also cannot

[146] See Yuk Hui's *Recursivity and Contingency* (2019).

gain access to the pre-individual charge of the ground of nature and humans.[147] If our world becomes proliferated with AI automatons as many anticipate, people will increasingly suffer from technological alienation. Such a world would appear to fulfill Heidegger's prophecy of "an absolutely technological state" brought about by the world movement of cybernetics (See Chapter 2). Nevertheless, if we adopt Simondon's theory of figure-and-ground in recurrent causality, subverting this technological alienation would be possible if AI research can move forward in a way that establishes a channel between the technical figure of AI and the ground of natural, psychic, and social milieux.

In the last two pages of *MEOT*, Simondon outlines how concretization can re-establish the rapport between technics and the metastable milieu of physical nature, human psyches, and collectives of transindividual relations, but the passage is brief and vague. Nevertheless, both Andrew Feenberg and Bernard Stiegler have respectively appropriated and further developed Simondon's theories of concretization and individuation in their social theories and their critiques of technology. Their attempts to turn Simondon's theories into critical social theories will be the focus of the next chapter.

[147] This also lends further support to my argument in Chapter 6 that deep learning AI is incapable of creativity or emotivity.

Chapter 8.

The Open and Closed Character of Concretization

Me: Is concretization a theory of openness or a theory of closed system?

ChatGPT: Concretization, as developed by Simondon, can be seen as a theory of openness rather than a theory of closed system. Concretization refers to the process by which a technical object becomes concrete, that is, the process by which it is transformed from an abstract, potential state to a concrete, actual state. This process involves the integration of various technical, social, and cultural factors into a coherent system.

Simondon's theory of concretization emphasizes the importance of understanding the technical object as an open system that is shaped by a variety of factors beyond its initial design or function. The process of concretization involves the integration of these factors into a coherent system, which can then be further developed and adapted to new contexts and situations. This openness to adaptation and development is an essential aspect of Simondon's theory of concretization.

Furthermore, Simondon's concept of concretization is closely tied to his broader philosophical framework, which emphasizes the importance of understanding the world as a process of becoming rather than a fixed, closed system. Simondon's philosophy emphasizes the idea that entities in the world are not pre-existing but instead emerge through a process of individuation, which is an open, dynamic process that allows for the emergence of new entities and new possibilities.

8.1. Introduction

In the previous chapter, I explained the physical basis behind Simondon's concept of pre-individual reality, and how this leads to his portrayal of the relations between beings in a figure-and-ground paradigm, in which the ground is the pre-individual reality and the source of potentiality for organic growth. Technological alienation is a consequence of the short-circuiting between the figure and the ground. As I will further explain in this chapter, concretization and transindividuation are the schemes to overcome this short-circuiting in technological alienation, as Simondon hints at toward the end of *MEOT*.

The idea of concretizing the technical and the social has been raised by both Bernard Stiegler and Andrew Feenberg, but they have seemingly come to the opposite conclusions. Influenced by Heidegger's dystopian critique of cybernetics, Stiegler argues that a social world concretized into the global computational and information system would turn into the resources, the standing-reserves, of a closed technical system. Feenberg, on the other hand, identifies the potentiality of transcending incumbent contradictions and stagnations when the social and the technical undergo the transductive operation of concretization. This identification matches Simondon's belief that concretization can overcome technological alienation and can possibly address Steigler's concern about the short-circuiting of transindividuation in automatic societies. As I will argue, these two views are analogous to two different perspectives of how a technical object evolves: the tendency toward robustness and stability after concretization, and the creativity in designs that transcend internal contradictions between modules and milieux. Stiegler's interpretation of Simondon simply points out that channeling between social critiques and sociotechnical innovation is a necessary condition for the transductive operation of concretization, and that this channel is being undermined by algorithmic governmentality. This forward-looking perspective stands in contrast to Feenberg's view on the past regarding how social movements have strengthened this channel of critiques. In this regard, they hold a consistent interpretation of Simondon's theories.

In the following, I begin by comparing the charge of pre-individual reality in Simondon to the Hegelian notion of potentiality that is influential in the critical theory of Marcuse and Feenberg. This parallel lends support to Feenberg's appropriation of Simondon's philosophy in his critical theory of technology. I then describe Feenberg's attempt to creatively apply Simondon's concept of concretization to his technical politics of resistant social movement. I argue that this attempt fulfills Simondon's original intention for concretization to overcome the chasm between culture and technics, but he only briefly sketches out this scheme toward the end of *MEOT*. For both Feenberg and Simondon, concretization is associated with a character of openness, as technology can evolve by virtue of inventors' creative imagination that follows a transductive logic, as opposed to incremental enhancements that follow deductive logic, or to the inductive logic in machine learning or any pattern recognition algorithm. I then explain Stiegler's re-interpretation of Simondon's concretization that takes into account the recent development of algorithmic governmentality and 24/7 capitalism. I will explain that Stiegler already identifies a closed character of technical concretization in his early writing, on how concretization leads to industrial standardization that cuts down the range of technical elements. In his later critique of automatic society and algorithmic governmentality, he considers the possibility of concretizing the technical with the human and the social, just like Feenberg, but his writing emphasizes the closed character of concretization under the totalitarian character of algorithmic governmentality. In his re-interpretation of Simondon's concepts, social institutions and their tertiary retentional systems are like the metastable supersaturated solution in crystallization, the introduction of new technology acts like the crystalline germ, and the psychic and collective individuation is analogous to the process of crystallization. As the transindividuation in algorithmic governmentality is automatized, psychic individuation no longer participates in collective individuation. The operation of critique and production of knowledge, an operation that corresponds to the open character of concretization emphasized by Feenberg, would be eliminated. I will conclude by explaining the dual character of concretization as pathways to both openness and closedness in technical lineages, and how the different emphases in Feenberg and in Stiegler reflect their different opinions on the extent to which the channel of social critique is bolstered by the technical politics of resistant movement or undermined by algorithmic governmentality and 24/7 capitalism.

8.2. Simondon's Pre-individuality and Hegel's Potentiality

Feenberg studied philosophy under Herbert Marcuse and is deeply influenced by Marcuse's philosophy of praxis. But he is also enamoured with Simondon's philosophy and has extensively referred to the theories of concretization and individuation in his own critical theory of technology.[148] He finds Simondon's philosophy compatible with his Marcusian critical theory, as both emphasize that philosophy should be concerned with practical human activities, that things and humans are fundamentally relational, and that

[148] For example, see *Transforming Technology* (Feenberg, 2002, pp. 186–188), and *Technosystem* (Feenberg, 2017a, pp. 66–85). Marcuse also draws from Simondon in his proposal of a "new science" (Marcuse, 1964, p. 237).

there is a potentiality for changes in the conflictual relations within a thing and between the thing and its environment. As I will explain in this section, this compatibility originates from the parallels between the Simondonian concept of pre-individual charge and the Hegelian notion of potentiality from which Marcuse developed his “two-dimensional” ontology (Feenberg, 2023, p. 84).

In philosophy, the term “potentiality” generally refers to any "possibility" that a thing can be said to have, whereas “actuality” refers to a possibility becoming real in the fullest sense (Durrant, 2015). Simondon rarely uses the term “potentiality” in *ILNFI.* Instead, he frequently employs scientific terms such as “potential energy” or “charge.” Thus, when he talks about the potentials associated with the pre-individual reality, he uses the phrase “charge of pre-individual reality” or “pre-individual charge,” which conjures up the electric charges in a battery, electrodes, or capacitors. From the following passage, it appears that his choice of words is made with careful consideration to distinguish potentiality as a simple possibility from the “real” potential energy in sciences:

> Gestalt theory attributes to the totality *simultaneously* the characteristics of a field and those of an organism; however, *the field exists before form-taking, and the organism exists afterwards*. Form-taking, envisioned as an operation of transductively propagated modulation, makes the real pass from the metastable state to the stable state and replaces a field configuration with an organism configuration. As a corollary, the energetic theory, such as we present it, of the operation of form-taking does not employ the notion of virtuality that is presupposed by the concept of good form; the potential, conceived as a potential energy, is real, for it expresses the reality of a metastable state and its energetic situation. Potentiality is not a simple possibility; it is not reduced to a virtuality, which is less than being and existence. (Simondon, 1964/2020, p. 710)

For Aristotle, “a thing that exists potentially does not exist in actuality, but the potential does exist” (“Actuality and Potentiality in Aristotle’s Philosophy | UPSC Notes,” n.d.). By distinguishing between potentiality and actuality, potentiality for Aristotle is a virtual possibility. Simondon argues that the good form in Gestalt theory, while it “attributes to the totality simultaneously the characteristics of a field and those of an organism,” also presupposes the notion of virtuality because “the field exists before form-taking, and the organism exists afterwards.” Whereas the good form in Gestalt theory is the stabilized and fixed form, the good form for Simondon is charged with energetic potential rich in energetic potential (Garelli, 1964/2020, p. xxiii), "charged with potentials actually existing as potentials, i.e., as an energy of a metastable system" (1964/2005, p. 352).

In his latest book on Marcuse, *The Ruthless Critique of Everything Existing* (2023), Andrew Feenberg also deliberates on potentiality as real possibility as opposed to formal possibility by referring to Kant, Hegel, and Aristotle. He alludes to the contrast between real possibility with formal or logical possibility in Immanuel Kant's *Critique of Pure Reason*. Kant's formal possibility refers to the fact that "[t]he human mind can construct imaginary objects that are incompatible with possible experience. Such objects are merely formal possibilities that cannot be realized" (Feenberg, 2023, p. 81). In contrast, "[r]eal possibility refers to possible objects that conform with the essential properties of experience. This type of possibility is a subset of the infinite variety of imaginable entities" (2023, p. 81). Whereas Kant's concept of real possibility has its basis in experience, Hegel appropriates this concept such that "it is no longer defined by the properties of experience in general. Instead, real possibility now relates to the logic of the thing itself of which it is the possibility" (2023, p. 81). Hegel got the idea of "the logic of the thing itself" from Aristotle's notion of potentiality. As Feenberg explains, "[i]t was Aristotle who first proposed a notion of potentiality. Substances have an essence which persists through change. This essence inhabits the substance and organizes it in a coherent whole" (2023, p. 85). Beneath the changes and developmental growth of a thing lies an unchanging essence or form that maintains its organization and coherence. This essence in substance models after living things: "Aristotle's model of potentiality was life. Living things realize a potential contained within themselves as they act and develop" (2023, p. 85). Aristotle's notion of potentiality assumes that the thing is substantial, that it is "a self-contained 'substance' with an inner essence that is only accidentally related to its appearance and other things" (2023, p. 87). Hegel "saves Aristotle's central idea: potentiality is not an extrinsic goal imposed on the thing but belongs to the nature of the thing itself" (2023, p. 87), but he overthrows the Aristotelian substance and his "version of essence [that] maintains the thing as what it is" (2023, p. 87). In Hegel's conceptualization of potentiality, "things are not, they become, and they do so through the interaction of their appearances and their environment, their inner and outer relations" (2023, p. 87). He underscores the conflicts and tensions of organisms and their milieux "that allows for the unfolding of essence, the development of potentiality into the actuality" (2023, p. 87).

To Feenberg, Hegel's notion of potentiality is "the basis for Marcuse's interpretation of the historical dialectic" (2023, p. 81). In Marcuse as well as in Hegel, "[t]he environment is not a harmonious resting place for life but a scene of conflict and struggle. But the fact of struggle is not ultimate. In overcoming the challenge of development, life absorbs the environment and the associated antagonism into itself. This is the process of realizing potentialities" (2023, p. 87). Marcuse sees two dimensions in a thing: a first dimension of its empirically given form and a second dimension of potentiality (2023, p. 85). His works "transcends philosophy in social theory" (2023, p. 88) by identifying this second dimension of potentiality as that which "transcends the given and opens the world to dialectical comprehension and revolutionary transformation" (2023, p. 84).

With this understanding of Hegel's notion of potentiality, we can see its similarity with Simondon's philosophy. There are parallels between Hegel's dissolution of independent Aristotelian essence in substance that maintains the thing as what it is, and Simondon's rejection of the form-matter paradigm in Aristotle's hylomorphism. Crystallization and sea anemones exhibit a mode of existence that cannot be associated with an unchanging form or essence. The becoming in individuation and concretization is thoroughly relational, as is the unfolding of essence in Hegel where the overcoming of conflicts and tensions in inner and outer relations allows the actualization of potentiality. The tensions and conflicts in the charge of pre-individual reality corresponds to the dialectic tensions and contradictions in Hegel's potentiality. And just as Hegel's notion of potentiality is taken up by Marcuse in the latter's social theory, Simondon's theory of individuation and concretization can also be appropriated to the social realm. As I will show in the remainder of this chapter, attempts were indeed made by Feenberg to adopt Simondon's philosophy in his technical politics, and by Bernard Stiegler in his analysis of algorithmic governmentality.[149]

[149] While there are similarities between Simondon's concept of pre-individuality and Hegel's notion of potentiality, Simondon pinpoints the difference between his concepts of individuation and transduction from Hegel's dialectics in *ILNFI*: "In this research, [Transduction] is called upon to play a role that dialectics could not play, for the study of the operation of individuation does not seem to correspond to the appearance of the negative as a second stage, but to an immanence of the negative within the initial condition through the ambivalent form of tension and incompatibility" (1964/2020, pp. 14–15 emphasis in original). Thus the key difference between individuation and dialectics is the "immanence of the negative" in the former and "the appearance of the negative as a second stage" in the latter. He then reiterates, "just like dialectics,

8.3. Concretizing the Social in Simondon and Feenberg

As explained in the last chapter, concretization is a theory about technical lineage in its most straightforward sense, but at a deeper level of understanding, it is a theory about how heterogeneous systems arrive at a state of coherence by establishing recurrences of causality between them. In this deeper philosophical sense, concretization can be a theory about a techno-geographic system as in the case of the Guimbal engine, which can be viewed as a technical system associated with the natural milieu of the oil and water turbulence. Concretization is an operation that can re-establish the rapport between technics and the geographic milieu, the former being the figures and the latter being the ground.

If the ground can be the geographic milieu under a recurrence of causality of concretization, can the ground also be the social milieu? Simondon did not elaborate much about the possibility of technical objects concretizing with the social world. But he did give a very abstract and vague picture of this possibility. On the last two pages of *MEOT*, Simondon discusses technical concretization as a "genetic method" that can be applied to "the study of the situation and role of technical thought in the whole [*l'ensemble*] of thought" (1958/2016, p. 247). This is because "technical objects cannot be considered as absolute realities and as existing by themselves" but "[t]heir technicity can be understood only through the integration of the activity of a human user or the functioning of a technical ensemble" (1958/2016, p. 246). Simondon never elaborates on how concretization can be the "genetic method" for how human activities can be concretized with technical inventions. As Andrew Feenberg and Gilbert Hottois have remarked, Simondon "remained vague on the political implications of his argument" (Feenberg 2016). But we can actually find such elaboration in the chapter "Concretizing Simondon and Constructivism" in *Technosystem* (Feenberg, 2017a, pp. 66–85). This book chapter attempts to reformulate "Marcuse's aspiration for harmony between human

transduction conserves and integrates the opposed aspects; unlike dialectics, transduction does not suppose the existence of a preliminary time as the framework in which the genesis unfurls, since time itself is a solution, a dimension of the discovered systematic: *time emerges from the pre-individual just like the other dimensions according to which individuation effectuates itself*" (1964/2020, pp. 15–16 emphasis in original). As Simondon further elaborates, "the synthesis [in dialectics] more or less envelops the thesis and antithesis by *overcoming* contradiction; the synthesis is therefore *hierarchically, logically, and ontologically* superior to the terms it joins together. Conversely, the relation obtained at the end of a rigorous transduction maintains the characteristic asymmetry of the terms" (1964/2020, p. 111 emphasis in original).

beings and nature … in a more empirically concrete form with the help of Simondon's theory of concretization" (Feenberg, 2017a, p. 69). By "concretizing" two seemingly incompatible theories, that of science and technology studies (STS) and that of Simondon's concretization, Feenberg paints a clearer picture of how technology can be integrated, not only to geographical nature, as Simondon illustrates in *MEOT*, but also to human nature:

> The theory of concretization explains how human and environmental contexts understood as associated milieux can be incorporated into design without loss of efficiency. This is not an outcome dictated by technological imperatives, but concretizing designs can in principle take account of these contexts as they do many others. Technology can thus be integrated to nature and to human nature. Struggles for environmentally sound technology, free expression on the Internet, and work that is humane, democratic, and safe are not extrinsic impositions on a pure technical essence but respond to the tendency of technical development to innovate synergisms of natural, human, and technical dimensions. They reveal the potentials awaiting realization. (Feenberg, 2017a, pp. 83–84)

According to Feenberg, the concretization of technology with the natural environment and human nature has been evident in the technical politics in the past few decades. The positive development in safeguarding our natural environment and individuals' rights follows the tendency to synergize the seemingly conflicting demands between the natural, human, and technical dimensions. It disproves that technical efficiency is necessarily in opposition to the needs of nature and of humans. This synergetic tendency follows the same principle as Simondon's concretization, as the "potentials awaiting realization" in a particular human or social milieu are transduced through its encounter with some emergent technical dimensions, resulting in a new sociotechnical ensemble. In this sense, by concretizing constructivism with Simondon, Feenberg has perhaps completed what Simondon ultimately wanted to achieve in *ILNFI* and *MEOT*, that is, philosophizing the "veritable complementary relationship" between man and machine (Simondon, 1964/2020, p. 425). The concretization that transduces man and machine into complementary relationship follows the same paradigm as the transduction between wave and corpuscle or that between the discontinuous crystalline lattice and continuous metastable crystalline solution.

In *MEOT*, Simondon tries to illustrate his vision of man-machine relationship with the supervisory role of a technician in regulating and looking after "the relation of the machine with the elements and the ensemble" (1958/2016, p. 78). But if we look at the entirety of his works, the technician example does not do justice to the inventive aspect of human participation and the resulting change in structures. In contrast, the democratic participation in technical evolution, from environmental social movement to the role of hackers in early history of computing and in free software movement, are much better illustrations of how the re-establishment of channels between figures (technology) and ground (nature, the psychic, and the social milieu) can liberate our society from technological alienation. They illustrate what Simondon calls transindividual relationship:

> [A]bove the social community of work and beyond the inter-individual relationship not supported by an operational activity, a mental and practical universe of technicity establishes itself, in which human beings communicate through what they invent. The technical object taken according to its essence, which is to say the technical object insofar as it has been invented, thought and willed, and taken up [assumé] by a human subject, becomes the medium [le support] and symbol of this relationship, which we would like to name transindividual. (1958/2016, p. 252)

A transindividual relationship refers to an intersubjectivity in the mental and practical universe of technicity. The most obvious example of transindividuality is associated with language and communication, which Simondon discusses in *ILNFI*:

> When the original charge of nature borne by individual beings cannot be structured and organized, there can be no form in the being for accommodating the form contributed by signals. To receive an information is in fact for the subject to carry out within itself an individuation that creates the collective rapport with the being from which the signal arises. To discover the signification of the message that stems from one being or several beings is to form a collective with them and individuate through the group individuation with them. There is no difference between discovering a signification and existing collectively with the being relative to which the signification is discovered, since signification is not of the being but between beings, or rather across beings: it is transindividual. (1964/2020, p. 344)

In this passage, Simondon explains how signals and information are meaningless by themselves. Language, on its own, is only an “an instrument of expression, the conveyance of information, but it does not create significations. Signification is a rapport of beings; signification is relational, collective, transindividual …” (1964/2020, p. 345). Signification, or meaning, emerges when the receiver of signals undergoes an individuation that put her or him in relation with the sender. This rapport across beings is a transindividual relation. In *MEOT*, Simondon generalizes transindividual relationship to the intersubjectivity in the mental and practical universe of technicity. Just like language, intersubjective relations are established over the invention and adoption of a technical object as the means to communication between human beings. Under this model of transindividuality, "[a]n inter-human relation … is thus created through the intermediary of the technical object" (1958/2016, p. 254). A human-technology relation is never one between a single individual with the technical object, but between a community via the technical object:

> The relation with the technical object cannot become adequate individual by individual, except in very rare and isolated cases; it can establish itself only to the extent that it will succeed in making this inter-individual collective reality, which we name transindividual, exist, because it creates a coupling between the inventive and organizational capacities of several subjects. (1958/2016, pp. 257–258)

What distinguishes a transindividual relationship from a technocratic society, in which forms of technology act as instruments of social control, is the channeling of energy from the natural and social milieu that plays a regulatory role on the evolution of these forms. While the voices of nature and the voices of human subjects are muted in a technocratic society, they are the sources of inventive schema in the individuation of technical objects and transindividual relationship.

8.4. Concretizing the Social in Stiegler

In the first volume of *Technics and Time* (1998) and in *Automatic Society* (2016), Bernard Stiegler also examines how the technical milieu may be concretized with the social milieu, but his conclusion is much less optimistic than Feenberg's analysis. Stiegler identifies the contemporary epoch as the latest stage of proletarianization in hypercapitalism. In this stage, the grammatization of Dasein's temporality and protention makes possible the industrial reproduction of human behavior, which is concretized into a sociotechnical milieu governed predominantly by algorithms. Here, Simondon's concretization is employed as a process that exacerbates the alienation of humanity, whose psyches and mental faculties are subjected to the control of algorithmic governance and surveillance capitalism. Yet, following Heidegger's eschatology in which salvation may come via an epochal change of collective awareness about the true condition of humanity (see Chapter 2), Stiegler perceives technology as primordially pharmacological, as both poison and cure, and the thinkers of our age are called to identify its poisonous character and participate in the invention of cure from within the realm of technology.[150]

In *Nanjing Lectures (2016-2019)* (2020), Stiegler reflects on his transition of thoughts on Simondon's concretization from his early writing on *Technics and Time* to his late writing on *Automatic Society*.

> When in the first volume of *Technics and Time* I discussed *On the Mode of Existence of Technical Objects*, I tried to show that the concretization of the industrial machine is effected (as the combustion engine, the electric locomotive or the turbine of the tidal power plant) when the latter must leave the purely technical milieu in order to form, with the natural milieu, a 'techno-geographical associated milieu', generated by the object itself, in the course of what Simondon calls its 'naturalization'. In *Automatic Society, Volume 1*, however, I argue that it is precisely in becoming a human (and not only physical) techno-geographical associated milieu … that this commodified retentional milieu, perpetually provoking, activating and calculating arrangements of retentions and protentions, leads to the psychic and social disintegration of retentions and protentions. (2020, pp. 316–317)

According to this passage, the first volume of *Technics and Time* explains the forming of a 'techno-geographical associated milieu' as a necessary operation in the concretization of the industrial machine, without contemplating the possibility of an associated milieu that is human or social. But Stiegler does in fact deliberate on the social implication of technical evolution (which can be interpreted as technical concretization) by identifying the power of technical system over other systems in industrial society:

> Industrialization is the affirmation of technological necessity. It is the sign of the immense power of the technical object over industrial society, of technical evolution in general over becoming in general, of the "technical system" over the "other systems." "At the industrial level, . . . the system of wants is less coherent than the system of the object; wants are formed

[150] Stiegler indicates that the "elaboration of a new *epistemology*, a new *philosophy*, and a new *organology*, in turn elaborating *a reconceptualization and a transformation of the digital* as such and on another basis than that of the computational ideology that took hold after the Second World War … will be the subject of *The Future of Knowledge*, the second volume of *Automatic Society* [italics in original]" (2016, pp. 33–34). Unfortunately, Stiegler did not live long enough to work on and complete this second volume.

> around the industrial technical object, *which thereby takes on the power of modeling a civilization*" (Simondon 1958, 24, my emphasis) … The technical system, the universal tendency that it carries, are no longer the partners of the "other systems"; the technical object lays down the law that is its own, it affirms an autonomy with regard to which, in the industrial age, the other layers of society must regulate themselves, with an actual possibility of negotiation. The indetermination of uses may well leave open possibilities for adjustments to the "system of objects," but at bottom the object bestows the horizon of all possibilities, essentially preceding the fixation of uses. (Stiegler, 1998, pp. 73–74 emphasis in original)

The "immense power of the technical object over industrial society," the "technical evolution in general over becoming in general," and "the 'technical system' over the 'other systems'" echoes Heidegger's premise in his technological will to will (see Chapter 2). Stiegler quotes from *MEOT* to affirm that technical evolution is dictated by technological necessity rather than human wants in Simondon's thoughts. It follows that technical object or system "lays down the law that is its own" and assumes "an autonomy with regard to which … the other layers of society must regulate themselves." Even though technology can be implemented with an indeterminacy that permits negotiations, the indeterminacy is limited, and in the end, technology "bestows the horizon of all possibilities, essentially preceding the fixation of uses." Stiegler attributes this closed character of technical objects to their concretization: "The concretization of technical objects, their unification, limits the number of their types: the concrete and convergent technical object is standardized. This tendency to standardization, to the production of more and more integrated types, makes industrialization possible, and not the converse …" (1998, p. 72). Concretization leads to standardization that limits the possible types of technical objects, reducing the level of indeterminacy that affords negotiations from the needs of other layers of society. This perspective of concretization seems contradictory to the open and inventive character of concretization that I allude to in the last section. Stiegler was in fact aware of this open character, though he interprets this as possible only in ruptures of successive epochs: "The dynamic play of the limit implies a discontinuity at the heart of all evolution in the sense of concretization. Ruptures mark the successive epochs in which the technical object gains its autonomy" (1998, p. 74).

The emphasis on the closed character of technical concretization is only vague and implicit in *Technics and Time*, but it becomes unambiguous and fully developed in his critique of computational society empowered by big data in *Automatic Society Vol. 1*.

Stiegler sees that "contemporary capitalism becomes purely computational, concretized in the so-called 'data economy'" (2016, p. 4), and our commodified sociotechnical milieu has become "a human techno-geographical associated milieu "via the digital exosomatic devices of that half of the world's population who are now equipped with smartphones, that is, personal portable computers, perpetually eliciting and capturing 'data', which is to say digital tertiary retentions" (2020, pp. 316–317). It is digital tertiary retention that absorbs the human into the technical system, with the human and the natural milieu both serving as an associated milieu in concretization: "With digital tertiary retention, techno-geographical milieus of a new kind arise, where it is the *human element* of geography that is associated with the becoming (that is, the individuation) of the technical milieu, so that this element itself acquires a technical function" (2016, p. 39, emphasis in original). Serving an indispensable role in a technical system, human individuals become a technical resource, a standing-reserve, for the apparatus of an industrial and commodified society:

> [W]hat gives rise to techno-geographical milieus in which psychic individuals become, through functional integration, functions of the apparatus of production and consumption is a process of the *concretization* (in the Simondonian sense) of the technical system that the 24/7 infrastructure itself forms qua functional integration of biological, psychic, sociological and technological automatisms. This forms what Simondon called an associated milieu, but in this case of a new type, and one he did not envisage. (2016, p. 80, emphasis in original)

Psychic individuals become "functions of the apparatus of production and consumption" under the "functional integration of biological, psychic, sociological and technological automatisms" of a technical system that has undergone a process of the concretizing the human and the natural milieux as associated milieux. Since Simondon did not envisage this new type of associated milieu in which human individuals are parts of the functional integration, Stiegler thus calls for "a reinterpretation of Simondon's thought with respect to contemporary realities" because "Simondon could be utilized in the service of the ideology of marketing, just as Foucault was used by liberals" (2016, p. 80).[151]

In his reinterpretation of Simondon's thoughts, Stiegler identifies the crux of the problem with the 24/7 capitalism in the short-circuiting of transindividuation, a concept

---

[151] By "ideology of marketing," Stiegler is referring to the "final integration of marketing and ideology" (2016, p. 66). This integration takes place "through the grammmatization of relations in which consists that traceability implemented by social networking" (2016, p. 66).

that comes from Simondon's formulation of transindividuality and of psychic and collective individuation, which I discussed in Section 8.3. The term "transindividuation" never appears in *ILNFI*, *MEOT*, or other Simondon's writing. Simondon only employs the terms transindividual (*transindividuel*) and transindividuality (*transindividualité*) to describe the intersubjectivity of psychic and collective individuation within the "mental and practical universe of technicity" (1958/2016, p. 252). As Yuk Hui explains, "Stiegler's concept of transindividuation refers to a transformation of the structures that compose the I and the We through the re-organization of tertiary retentions" (2014). Hence transindividuation can be understood as an alternative term for Simondon's psychic and collective individuation in which the transindividual (inter-individual) relationship is mediated by technical objects. But in Stiegler's appropriation of Simondon, the technical medium of interest is not the "mental and practical universe of technicity" that concerns Simondon, but digital tertiary retention. To see why this appropriation is significant to Stiegler's critique of the full and generalized automation of the 24/7 global computational infrastructure, we must first understand his pharmacology, which is a theory about the simultaneous toxic and curative nature of technology.

8.5. The Short-Circuiting of Transindividuation

Stiegler's pharmacology is captured in the concept of "doubly epolkhal doubling" (2016, p. 12). He develops this concept to describe

> how a shock begins by destroying established circuits of transindividuation, themselves emerging from a prior shock, and then gives rise to the generation of new circuits of transindividuation, which constitute new forms of knowledge arising from the previous shock. A techno-logical *epoché* is *what breaks with constituted automatisms*, with automatisms that have been socialized and *are capable of producing their own disautomatization* through appropriated *knowledge*: the suspension of socialized automatisms (which feeds stupidity in its many and varied forms) occurs when *new, asocial automatisms are set up*. A second moment of shock (the second redoubling) then produces new capacities for dis-automatization, that is, for negentropy to foster new social organizations. (2016, p. 12)

In this passage, Stiegler equates a technological epoch with certain established circuits of transindividuation, which emerge from two moments of shock. A first moment of shock is associated with "a new form of tertiary retention" (2016, p. 34), which destabilizes the constituted automatisms in the prior epoch. The technological shocks of alphabetic writing, the printing press, the Internet are examples of this first moment. In a second moment of shock, the new form of tertiary retention "requires the formation of *new knowledge*" in order to "achieve socialization, that is, collective individuation" (2016, p. 34). "New ways of doing things and reasons to do things, of living and thinking" are constituted through such new knowledge (2016, p. 34). This would result in "new forms of existence" and "new conditions for subsistence" (2016, p. 34), which would foster new social organizations. For instance, after the popularization of the Internet via the World Wide Web, which was initially used for informational retrieval, the second moment of shock came from the new knowledge that the Internet makes it possible to democratize the production and consumption of knowledge and information, thus destabilizing the purpose and the authority of centralized media.

The democratization of the early Internet demonstrates how technology as "the *pharmakon* can be toxic and curative only to the extent (and in the excess) that it is both entropic and negentropic" (2016, p. 100). But the recent application of big data and deep learning in 24/7 capitalism has been "a negentropic factor that has become massively entropic" (2016, p. 100). How can a second moment of shock come about such that this new *pharmakon* become more therapeutic rather than toxic? According to Stiegler, this

> must proceed from a social innovation that reinvents the adjustments between the social systems and the technical system, and does so according to a model where it is no longer the economic system, and the technological innovation it requires, that prescribe the social. It must on the contrary be a model in which social innovation, founded on different economics - on a contributory economy - and on a reinvention of politics, conceived as therapeutics, prescribes technological innovation, that is, organological evolution, and does so by interpreting technical tendencies.
>
> Such a becoming is highly improbable. Yet it alone is the bearer of the future, that is, of the rational. (2016, p. 100)

This analysis seems like a generalization of the therapeutic transformation in alphabetic writing, in the printing press, and in the democratizing movement of the early Internet. For such social innovation and reinvention of politics to be possible, a society must have an established mechanism for political individuation:

> [P]olitical individuation is founded on the collective critique of its sources, that is, on the critique of circuits of transindividuation that it inherits in the form of various forms of knowledge, themselves organologically constituted, concretized, *judged* and as such *realized* through institutions

> that are always retentional systems. These institutional sources, which metastabilize the transindividual, reconvert them for any psychic individual into the preindividual funds that its social milieu forms, charged with potentials for individuation. (2016, p. 145 emphasis in original)

Institutions, such as those made up of government legislations or collective bargaining agreements, are always retentional systems. These systems “metastabilize the transindividual” because they provide a stable ground for shared life-knowledge yet always subject to critiques and changes. This metastable ground, like the supersaturated crystalline solution, is the site for contention between conflicting parties that is the source of the preindividual funds and potentials for individuation. The transition into new tertiary retentional systems, much like the crystalline seed, should introduce new possibilities that bring about the actualization of potentials for individuation.

The problem with the tertiary retentional system of algorithmic governmentality is its elimination of the need to resolve conflicts though contentions and negotiations. As Stiegler argues, this elimination would lead to the destruction of “signification,” and along with it, of the possibility of the new knowledge necessary for a second moment of shock:

> Signification [*signification*], that is, *semiosis* as engendering signs, significations and significance (making-signs), is the transindividual made possible by the process of transindividuation woven between psychic systems, technical systems and social systems – that is, between psychic individuations, technical individuation and collective individuations.
>
> The destruction of signification by the digital technical system results from the technology of power deployed by the algorithmic governmentality of 24/7 capitalism, and it is founded on eliminating processes of disparation. The latter is a concept that Simondon introduces … (2016, p. 128)

This relation between signification and disparation can be clarified by reading *ILNFI*. As discussed in Section 8.3, signification for Simondon is “a rapport of beings; signification is relational, collective, transindividual” (1964/2020, p. 345). Language is one instrument that facilitates such transindividual relation. Simondon argues that signification “emerges from a disparation” and presupposes “the existence of a system in a state of metastable equilibrium” (Simondon, 1964/2020, p. 16).[152] Signification is the new dimension

[152] The discussion on disparation can be found in Section 6.4 and pre-individuality in Chapter 7.

introduced to resolve disparation, in which incompatibilities become positive constituents of the new dimension (See Section 6.3). Drawing from the *Wired* Article “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete” (C. Anderson, 2008), Stiegler raises the concern that the replacement of theoretical knowledge with automated knowledge in algorithmic governmentality eliminates the social process of knowledge creation, which involves negotiations among stakeholders with conflictual interests. In Simondon’s terms, automated knowledge eliminates the state of disparation in a metastable equilibrium from which signification emerges. It is like turning a metastable supersaturated solution into an unsaturated solution, which no longer holds the pre-individual potentiality for crystallization.

For Simondon, collective individuation can be viewed as the next layer of individuation on psychic individuation, as individuation that transduces the disparations and the tensions of pre-individuality unresolvable at the level of psychic individuation. The living being is a product of synthesis and unification not only of complete physical individuals, but also of physical beings whose individuation is incomplete.[153] Whereas a physical being will eventually exhaust the potential energy of its pre-individual reality, such as the completion of crystal formation in crystallization, the potential energy of a human being is not exhausted until it dies. While alive, the potentiality for its perpetual growth lies in the “conflict between the pre-individual reality and the individuated reality within the subject” (Simondon, 1964/2020, p. 354). Simondon suggests that this conflict corresponds to the inner contradictions and the irrational intensities beneath conscious mental activities. It

> indicates to the subject that it is more than the individuated being and that it contains the energy for a further individuation; but this further individuation cannot take place within the being of the subject; it can only take place through this being of the subject and through other beings as the transindividual collective. (Simondon, 1964/2020, p. 354)

A transindividual collective “is not a milieu for the individual but a set of participations in which it enters through this second individuation” (1964/2020, p. 348). It is “the signification obtained by the superposition of beings that are disparate by themselves in

---

[153] “[W]hat makes the living being appear is in a sense the suspension of the development of the physical being and its analysis, not a synthetic relation which unites completed physical individuals” (1964/2020, p. 168).

a single system" (1964/2020, p. 349). Accordingly, contradictions and irrational intensities are the sources for personal and social growth.

To shed light on the implication of disparation on psychic and collective individuation, Stiegler draws from the dilemma faced by individuals who belong to social groups with conflicting ideals:

> [T]he psychic individual is related to and individuates with collective individuals that may be in contradiction with each other. This is why it must project itself onto the plane of a transindividuation of reference in order to overcome these contradictions … and the superego is one such plane, which can be embodied in artificial crowds like the Church or the Army, but also in other forms, like the Fatherland, the Party, and so on. (Stiegler, 2016, p. 152)

By collective individuals, Stiegler is referring to artificial crowds like the Church, the Army, the Fatherland, or the Party. A psychic individual may belong to both the Church and the Army, which may be in contradiction with each other. To overcome such contradictions, the individual "must project itself onto the plane of a transindividuation of reference." Contradictions between artificial crowds are reorganized into positive constituents of this new plane, and such resolutions or transductions of contradictions bring about the unique development of an individual's character. Transindividuation thus refers to the co-individuation of a psychic individual and collective individuals, which are associated with social institutions and technological apparatus.

These contradictions, or disparations, are sources of the potentiality for transindividuation. Under algorithmic governmentality and 24/7 capitalism, the "process of automatized and fully computational transdividuation erases the potential for 'disparation'" by "dissolving collective individuals" (Stiegler, 2016, p. 152). In "algorithmic concretization[,] … noetic time can be outstripped" (2016, pp. 149–151). Stiegler defines "noetic individuals" as "intellectual and spiritual individuals" (2016, p. 12). Thus "noetic time" is the time for intellectual and spiritual reflection, which is crucial for the contemplation of social critiques. Societies prior to 24/7 computational infrastructure have built-in means and structures for facilitating dialogues over social critiques, which then feed back into the process social and technical innovation, gradually leading to the overcoming of disparations in transindividuation. But under algorithmic concretization,

> psychic individuation no longer participates in collective individuation and, as such, if we have read Simondon well, goes nowhere [*tourne à vide*]. … The process of transindividuation and the transindividual are … *automatized and concealed by the speed of their production*, and founded on this high speed. This outstripping [*prise de vitesse*] of psychic and collective individuals is a *taking-apart [déprise] of the form of the noetic* by the computational formlessness [*informe*] in which this speed consists. (2016, p. 149, emphasis in original).

As the form of the noetic is taken apart, an individual is no longer an indivisible unit, but divided into digital parameters in computational algorithms. Stiegler coins the term trans-dividuation to denote "this false transindividuation that short-circuits psychic individuals" and to contend that "[t]he process of transindividuation and the transindividual are replaced by the transdividual and transdividuation" (2016, p. 149). An electrical circuit is "short-circuited" by connecting two points on the circuit with a wire, and electric current will pass through this wire instead of through the normal circuit. When Stiegler uses this metaphor, the "wire" represents the algorithmic and computational infrastructure. The transductive operation of incorporating psychic individuals into social collectives is bypassed by this "wire," as the individuals now exist in the mode of digital parameters that are algorithmically integrated with the computational infrastructure of 24/7 capitalism.

Much like Feenberg, Stiegler extends the theory of concretization to include the social milieu as an associated milieu of a technical object, but their conclusions seem vastly different. Whereas Feenberg and Simondon identify the potentiality in concretizing the social, Stiegler points to the danger therein, as demonstrated by the algorithmic and automated infrastructure in our computerized society.

8.6. The Open and Closed Character of Concretization

We can gain some clarity on the seemingly contradictory stance between Feenberg and Stiegler on their appropriation of Simondon's concretization by pointing out the simultaneously open and closed character of concretization in technical evolution. Recall that technology evolves when inventors creatively imagine "an invention that *presupposes the problem to be resolved*" (Simondon, 1958/2016, p. 57 emphasis in original). This creative imagination cannot be arrived at by deductive or inductive thoughts, but only by transductive thoughts. Technology is always open to evolve because it shares its pre-individual reality with the natural milieu and with the human milieu, which come with conflicts, contradictions, tensions. Simondon calls them disparations, which hold the potential energy for changes. This openness of technical evolution typically involves the branching from technical lineages into new lineages in the overcoming of disparations. At the same time, within a single technical lineage (such as the examples of diodes in *MEOT*), the becoming more concrete implies a progress toward perfecting the relations between technical organs (modules) to achieve internal coherence, thereby increasing the robustness of the technology and reducing its margin of indeterminacy. From this angle, as a technology becomes more concrete, as its design becomes more perfect, the room for enhancing it also diminishes correspondingly. Thus, technology tends toward becoming a closed system in the process of concretization if there is no change in the natural or human milieu that compels a restructuring of the technology.

We can sense the closed system tendency of concretization in Stiegler's remark that "algorithmic governmentality and 24/7 capitalism are the worldwide and total concretization" (Stiegler, 2016, p. 121). Algorithmic governmentality is concretized in a recurrent causality where its disruption of the social process of actualizing revolutionary potentiality in technical politics contribute to a favorable environment for algorithmic probability to subsist. If the process of disparation is precluded in algorithmic governmentality and 24/7 capitalism, the pre-individual charge of individuals and collective would be inhibited from transducing changes in the system. From this perspective, the open character of concretization is inhibited unless a second moment of shock brings about the re-establishment of new circuits of transindividuation, which permits the flow of potential charge from the pre-individual reality of the human and natural realms. It is fair to say that Stiegler's reinterpretation of Simondon's thoughts remain true to the logic behind the theories of concretization, transduction, individuation, and transindividuation, and at the same time, it puts into question the Simondonian idea that concretization can overcome "the opposition drawn between culture and technics, between man and machine," an opposition that "is false and has no foundation" (Simondon, 1958/2016, p. 15).

This overcoming of the opposition between culture and technics, roughly sketched out by Simondon toward the end of *MEOT*, can indeed be actualized in a technopolitical environment that favors grassroot resistance. Feenberg emphasizes the perspective that, despite the proliferation of automated technologies, human agencies today are still active participants in the operations and inventions of contemporary technology. The Internet has inherited many characteristics from the preceding, more democratic, Internet milieu in the 1990s, in which grassroots and marginals could increasingly participate in technical decisions and political actions. The increase in grassroot participation in collective decisions has helped re-establish the recurrence of causality between the technics and the social, mending the broken channel between the figures and the ground of pre-individual charge. If we agree that today's Internet encompasses a mixture of technocratic and democratic aspects, it seems more correct to characterize the current mixture as another metastable milieu filled with the tensions of pre-individual reality, from which new inventions will eventually emerge, propagating the individuations in psyches and social collectives.

Unlike Feenberg, Stiegler focuses on how the channel between social critiques and technical development has been undermined by algorithmic governmentality and 24/7 capitalism. Over the past couple of decades, we have indeed witnessed a diminishing room for democratic activities on the Internet under surveillance capitalism, with power elites gaining control over people's desire and behavior. Such algorithmic governance can be modelled as cybernetic feedbacks, with a clear objective of recursively narrowing the gap between ideal behavior (e.g., purchasing a product) and actual behavior. If this trend continues, the increasing infiltration of AI technologies in everyday life will make our society suffocate from conflictual expressions and purify it into a smooth and perfect human-machine coupling. The alarming critique of Stiegler could perhaps be applicable if his future projection turns out to be correct. In short, whereas Feenberg looks at the past and points out the improvement of this channel from social movement and technical politics since the mid-twentieth century, Stiegler looks toward the future and anticipates the progressive elimination of noetic time due to algorithmic governmentality and 24/7 capitalism. Such elimination makes people incapable of critical reflections necessary for concretization and transindividuation.

8.7. Conclusion

This chapter presents how Feenberg and Stiegler respectively appropriated Simondon's theories of concretization and individuation to formulate their critiques of technology. Feenberg finds Simondon's philosophy generally compatible with his Marcusian critical theory. Marcuse's two-dimensional ontology is based on Hegelian potentiality, and both Hegelian potentiality and Simondon's pre-individuality identify the potential for changes in the conflictual relations within a thing and between the thing and its milieu. And just as Marcuse finds the basis of his social theory in Hegel's notion of potentiality, Feenberg adopts Simondon's theory of concretization in his technical politics. Concretization is a transductive operation that brings about a techno-geographic system in which the technical system is associated with a natural-geographic milieu in a relation of recurrent causality. Both Feenberg and Stiegler appropriate concretization by substituting this natural-geographic milieu with the social milieu as the associated milieu of the technical system, but they have seemingly come to opposite conclusions on the openness of the system. Feenberg identifies the openness of concretization in transcending the tensions and contradictions of incumbent sociotechnical systems and finds actual realization of this in the political struggles for environmentally sound technology, free expression on the Internet, and work that is humane, democratic. Stiegler identifies the closed character of a technical system in concretization, as a technical system is less malleable to changes as it tends toward perfection and concreteness. When the human and the social is absorbed into this concretizing system, they become resources and standing-reserves for the expansion of realms under technical controls. Whereas Heidegger sees cybernetics as the culmination of the technological will to will, Stiegler sees algorithmic governmentality and the associated tertiary retentional system as totalizing in their manipulation of pretention in human behavior.

I then try to elucidate these two perspectives on concretizing the social by explaining the simultaneously open and closed character of concretization in technical evolution, which has an open trajectory across technical lineages but a tendency toward closedness within a single technical lineage. The branching off from one technical lineage to another demands the charge of pre-individual reality from psychic tensions and the social conflicts, and if the channeling of the charge to technical invention is blocked, the sociotechnical system becomes a closed system. In the end, the difference between Feenberg and Stiegler reflects their different opinions on the condition of the communication channel between social critiques and the process of inventing technologies. Feenberg emphasizes the improvement of this channel in the sociopolitical environments since the mid-twentieth century, whereas Stiegler raises concerns on the disconnect of a society from its critical voice under algorithmic governance. The only possibility of overcoming the tendency toward totalizing automatization must proceed from a social innovation that reinvents the adjustments between the social systems and the technical system. But when the ground, the charge of pre-individuality in the human psyche and social tensions, has been blocked from channeling the energy to the inventive cycles of technology, the sociotechnical system becomes resistant to changes.

There is a certain degree of speculation in Stiegler's prognosis about the future automatic society. Therefore, it is crucial for us is to evaluate this future prognosis by considering what I have laid out so far in the previous chapters on machine learning and cybernetics. In particular, Stiegler often conflates cybernetics with computing. As I have argued in Chapter 6, cybernetics is associated with a philosophy of openness, whereas it is the realm of computing that is associated with universality. Many historical inventions of computer-related technologies came out of the combination of ideas from cybernetics and computing, leading to the common mistake of conflating the two.[154] One of the cybernetic ideas that have been adopted by the computing industry is that of symbiosis (Bardini, 2000). Human-computer symbiosis was in fact the fundamental idea behind the invention of the windows operating system and the mouse in personal computing (Bardini, 2000), which grew out of the countercultural movement in the 70s (Turner, 2008). For many belonging to computing industry, including Kai-Fu Lee and Qiufan Chen who authored *AI 2041* (2021), human-computer symbiosis is often seen as the antithesis of AI automation. They see human-machine coupling as possibly liberating if the two parties are engaged in a symbiotic feedback relationship rather than a deterministic relationship dictated by formal logics. Simondon has also argued that human and machines ought to be in a relationship as peers rather than as master-and-slave. Will a symbiotic relationship between the human and AI be liberating for the human or will it further reify the human into a slave of AI? In the next and concluding chapter of this dissertation, I will deliberate this question by examining the stories and analyses in *AI 2041* while leveraging our understanding of AI and machine learning from the historical and philosophical analyses in the preceding chapters.

[154] See footnote 56.

Chapter 9.

A Critical Inquiry of Human-AI Symbiosis

Me how does human-computer symbiosis contrast with automation?

ChatGPT: Human-computer symbiosis and automation are two different approaches to using technology to improve productivity and efficiency in various tasks.

Automation refers to the process of using machines, software, and other technologies to perform tasks that were previously done by humans. The goal of automation is to eliminate human involvement in repetitive, low-skill tasks, and reduce the cost of labor. In this approach, the computer system or machine is responsible for performing the task, and humans are not directly involved in the process.

On the other hand, human-computer symbiosis is the idea that humans and computers can work together to achieve common goals. The focus of this approach is on enhancing human abilities rather than replacing them. The goal is to create a partnership between humans and computers, where each contributes their unique strengths and skills to solve problems and complete tasks. In this approach, computers are used as tools to augment human abilities and provide new ways of thinking about and approaching problems.

In contrast to automation, human-computer symbiosis emphasizes the importance of human involvement in the process. While automation seeks to replace human involvement in certain tasks, human-computer symbiosis recognizes that humans bring unique skills, creativity, and problem-solving abilities that cannot be replicated by machines alone. Instead, the goal is to create a synergy between humans and computers, where each can contribute to the task in their own way.

9.1. Overview

Up to this point in the dissertation, I have laid the groundwork for understanding machine learning by tracing its relation to cybernetics and by deliberating the philosophical critiques of cybernetics by Martin Heidegger and by Gilbert Simondon. Whereas Heidegger sees cybernetics as the culmination of the essence of modern technology, as the complete realization of the technological will to will, Simondon derives a philosophy of openness from cybernetics and suggests a possible trajectory of co-evolution between the human, the social, and machines. Stiegler and Feenberg have both taken up Simondon's philosophy in their analysis of contemporary technics. Feenberg appropriates the openness of Simondon's theory of concretization to explain the technical politics of social movements. Stiegler examines the closed nature of concretization and the short-circuiting of transindividuation to present a critique of algorithmic governmentality. Thus for Stiegler, the technical politics advocated by Feenberg may not be sustainable in an automatic society in which algorithms hold the position for key decision-making over the human minds.

But will developments in AI empowered by deep learning necessarily bring about an automatic society that numbs individuals' capacity for social critiques and thereby short-circuits the transindividuation of the human, technics, and society? Can there be alternatives for AI to co-evolve with the human and our society? Some computer scientists offer the research direction of human-computer symbiosis as one possible alternative to AI automation.[155] This direction, endorsed by the authors of *AI 2041* as well as Hubert Dreyfus[156], appears to be most compatible with Simondon's proposal that humans ought to relate to machines as peers rather than as masters and slaves. In fact, throughout the history of computing, symbiosis and automation represent two competing visions for computing research. Whereas automation, along with the cyborg imaginary, is associated with the reduction or elimination of human faculties, research on human-computer symbiosis explores how the human body and its faculties can be augmented via their interactive feedbacks in computers. But as I will contend in the following, the computer scientists' vision of human-AI symbiosis is concerned with how humans relate to technology at an individual basis but not with how the society may be transformed. Hence the vision can easily be co-opted in a technocratic society because it lacks the dimension of transcending incumbent sociotechnical contradictions, which as I have explained in Chapter 8, is implicit in Simondon's theories of concretization, individuation, and transindividual.

---

[155] Human-AI symbiosis is the subject of numerous articles in computer science journals (E.g., Becks & Weis, 2022; Jarrahi, 2018; Mahmud et al., 2022; Zahedi et al., 2022; Zahedi & Kambhampati, 2021; Zhang et al., 2022).

[156] See footnote 20.

9.2. A Critique of AI Automation

Critics of AI today, such as Stiegler whom we discussed in Chapter 8, are mostly concerned with the trend of exacerbating our overreliance on technical solutions that are opaque to human understanding and supplant the human in critical decision-making within the global informational infrastructure. Anticipating the trend's social implication brings out ethical questions from the loss of human dignity in mass job replacement to the existential risk of AI misalignment in which AI agents would be put into self-reflective loops and yield behavior misaligned with their designed purposes. As Yuk Hui explains, the exacerbation of such social issues can be traced to the life-like recursivity and contingency in computer operations and software processes, which behave like gigantic organism: "we are observing the becoming of an 'artificial earth,' and we are living within a gigantic cybernetic system" (Hui, 2019, sec. 6). This metaphor seems even more apt recently, with the way people are experimenting with automating large language models that yield curious behaviors (See *As AutoGPT Released, Should We Be Worried about AI?*, 2023). For instance, a program built on the OpenAI's GPT-4 API was given the problem of figuring out how to solve a website's CAPTCHA test, and it came up with a rather creative solution. The program went on the TaskRabbit platform and, over exchanges of text messages, it tricked an unsuspecting human worker into helping it solve the CAPTCHA test by pretending to be a blind person (Kan, 2023).

Living within a milieu of these gigantic cybernetic systems, humans would appear minute in scale, gradually forsaking their agency and freedom to cybernetic machines. The problem with the increasing marginalization of the living in favor of a cybernetic milieu lies in the fact that the organic appearance of AI technology belies its essence as a technical or scientific object, which as Feenberg argues, has no meaning if placed outside the context of human societies. In his analysis on objects constructed by sciences, Feenberg argues that scientific objects attain their meanings only in relations to the experiences of ordinary humans who live with the real-life version of those constructed objects (see Feenberg, 2023, Chapter 6). Since the products of AI-generated results are essentially objects constructed by sciences, Feenberg's argument can be appropriated for the necessary relation between AI and the human. No matter how complex is the behavior of an AI program, its results are like any object constructed by sciences and must be evaluated by human reasoning. The human intelligence must adjudicate AI-generated results with aspects of scientific rationality and teleological notions coming from the lifeworld perspective. This adjudication is an art that uses science to understand and correct the products of AI, which lacks the human capacity to evaluate its own results in a wider context. With the human taken out, historical contingencies, anomalies, and cultural specificities can invalidate any result computed by AI, and they could bring about the obsolescence or inappropriateness of contextualized data patterns that had been used to train AI models.

Therefore, just like other technical or scientific objects, AI technologies are meaningless objects outside the human context. Relying unthinkingly on AI is analogous to the overreliance on a smart camera to get a good photograph. Even with a smart camera, a photographer needs to take many photographs to find a good one, and only human intelligence can make the selection. In analogy, if we train an AI model with a bunch of Cezanne paintings, we will certainly succeed in making a program that is capable of recognizing Cezanne's style and of regenerating paintings of the same style. But what AI cannot do is to select aesthetically pleasing results out of thousands of regenerated paintings in Cezanne's style without human intervention, since what are considered aesthetically pleasing depend on a changing cultural context and political environment.[157] This is comparable to my argument in Chapter 6 that AI is not capable of composing music that react to political suppression like the anti-Stalin messages in Dmitri Shostakovich's symphonies.

The human-AI relation becomes problematic when humans falsely attribute intelligence to AI programs and consequently, relinquish their role in adjudicating AI-generated results. Such is the case with the employment of AI-assisted decision-making in law courts (see Zerilli et al., 2019). As Simon Natale (2021) has argued, deceitfulness is part and parcel of what makes AI appear intelligent throughout its historical development. User experience can be characterized as a leap of imagination beyond the

---

[157] This argument takes the view that what is aesthetically pleasing needs to be adjudicated with aspects of scientific rationality and teleological notions. Some may refute this view in the Cezanne example, arguing that what is aesthetic pleasing is a matter of taste, that *les goûts et les couleurs ne se discutent pas* (everyone's taste is different). With this view on taste being cultural relative, AI can be trained to be up to date on recognizing patterns that match the latest fashion or taste in a particular culture. But even then, this AI adjudicator of beauty needs to be trained based on human tastes and is simply an algorithmic representation of the taste of certain culture at a particular moment in time. It can never be completely automated and become the standard of beauty on its own.

strict functional specification of a software application. In the case of AI, this leap of imagination often becomes the false attribution of intelligence to a technical object. Consider the large language model. From the perspective of functional specification, the software program models human languages by representing the semiotic relations between words or sentences with the proximity of numerical vectors. This model is based on how linguists model human languages. It is designed to perform translations or to carry out conversations like humans, but not designed to reason like humans. Yet, when people engage in dialogues with ChatGPT, they are inclined to attribute the program with reasoning abilities to such an extent that Geoffrey Hinton foresees these language models attaining an IQ of 210 in the near future (Joseph Raczynski, 2023). But the functional specification of language modelling, whose primary objective is to imitate human dialogues, clearly has nothing to do with genuine human reasoning or the ability to understand and produce meanings.[158]

If the world approaches a state in which AI agents of such ilk are responsible for running all major societal processes and decisions, to the point where humanity becomes irrelevant to these processes that are more concerned with the technological will than with the goods of humanity, such an extreme case of AI automation would effectively be no different from an abandoned windmill. Purely automated AI may initially align with social values, but without human interventions their original alignment with human purposes would be invalidated by anomalies, historical contingencies, and cultural specificities. So eventually, fully automated AI would necessarily diverge from the contingent development of humanity. In the hypothetical scenario where humans become marginalized to the point of irrelevance or even extinction due to AI, the complex technical processes may keep on running like an abandoned windmill, but neither automated AI nor the windmill would fulfill any real purpose. In other words, a sociotechnical advancement toward this extreme case of AI automation is effectively a collective progression toward a dead world.

Could future societies be foolish enough to delegate every crucial decision to AI, buying into the deceiving appearance of creativity and of understanding human meanings? Such a scenario is certainly possible but not inevitable. At the current

[158] On the specific reasons behind why large-language models cannot understand meanings, see "Climbing towards NLU: On meaning, form, and understanding in the age of data" (Bender & Koller, 2020).

juncture, it is ambivalent whether societies will choose a path toward full AI automation in key decision-making.[159] Proponents may argue that AI automation can help mitigate the evilness and propensity to wars inherent in human societies. But if an overdependence on AI automation leaves no room for human decisions, we ought to make every effort to prevent the emergence of a world void of meaning by raising public awareness about this inherent lack in AI. Should we then condemn all efforts on AI automation? Have we not experienced the benefits brought about by the many hidden AI processes that partially automate tasks for our convenience? One alternative to full AI automation is the research on human-computer symbiosis conducted by J. C. R. Licklider and Douglas Engelbart in the late 1960s and early 1970s. In their vision, partially automated AI does not necessarily reduce or replace the human, but instead can augment human capabilities and expand the possibilities for how the human can live.

9.3. Automation versus Symbiosis

Automation and symbiosis characterize the two prevailing visions for computing research and development. Over the history of universal computing, researchers have explored different realizations of cybernetic ideas that may give either the sense of automated mechanization or the sense of symbiotic openness. According to Thierry Bardini,“[t]wo main projects had stemmed from the rise of cybernetics: intelligence amplification, including Engelbart’s Augmentation of the Human Intellect project, on the one hand, and the effort to produce artificial Intelligence [*sic*] (AI), on the other” (2000, p. 19). Whereas AI automation tends to simulate the human behavior in machines and to substitute the human in an information workflow, research on human-computer symbiosis, pioneered by Licklider and Engelbart, explores an open relationship of cybernetic feedbacks between the human and the computer for the purpose of augmenting human intelligence and other abilities.

The original idea of human-computer symbiosis can be traced to the seminal paper “Man-Computer Symbiosis” (1960) by Licklider, who provided funding support for the Augmentation Research Center at Stanford Research Institute. This research center

[159] Feenberg made a similar point about the ambivalence of computer technology (2002, pp. 89–130).

was headed by Engelbart, whose vision of the co-evolution of the human and the computer represents the center's research direction in the 1960s. It has been credited with the first inventions of the mouse and of the precursor to window-based graphical user interface.[160] In Chapter 6, I argued that cybernetics is associated with both the reductionism of the living as well as the theme of openness and complexity. In the context of computing technologies, the reductionism of cybernetics is manifested in AI automation whereas the openness and complexity of cybernetics are manifested in symbiotic human-computer relationship.

While technology can be reifying, prescribing a pre-determined set of actions for users interacting with the technology (Akrich, 2010), technology can also open a horizon for new possibilities of actions and creations. Language, writing, software programming language, networking protocols are all examples of technologies that both prescribe the grammar for actions and open a space for creative imagination. The degree of openness as well as the degree of behavioral inscription vary with the way a technology is designed.[161] The more open is a technology, the greater the potential for humans to evolve symbiotically with the technology. At the same time, this symbiotic engagement comes with the behavioral inscription encoded with values and biases, which need to be reflected upon and critiqued by the user community. In a symbiotic relation, the human and the open technology co-evolve in a dialectic movement between creative actions and reflective critiques.

This co-evolution is also possible in a cyborg. The term "cyborg," a portmanteau of cybernetic and organism, was coined in 1960 by Manfred Clynes and Nathan S. Kline (1960). A cyborg is a being with both organic and biomechatronic body parts, whose functions are performed by software and chips embedded in computerized devices. The cyborg imaginary is similar to human-computer symbiosis in that both depend on a communication interface between the human and the machine. As in a symbiotic relation with a computer, a person whose vision is enhanced by an artificial eye may be opened to a new horizon of possible actions. Nevertheless, there is a subtle difference between

[160] Thierry Bardini documents the history of such inventions in *Bootstrapping: Douglas Engelbart, Coevolution, and the Origins of Personal Computing* (2000).

[161] I elaborated on this dialectic of open technology in my previous works (Lo, 2016, 2019).

the cyborg imaginary and human-computer symbiosis.[162] Whereas the cyborg imaginary focuses on the assimilation and potential replacement of human faculties, Engelbart regards the human and the computer as symbiotic partners that co-evolve over their interactions, and he modelled such interactions as cybernetic feedback loops between users and machines.[163] In a perpetual cycle, users in symbiotic relations with computers would incessantly adjust and augment their bodies, minds, and senses in response to an evolving milieu of technologies. At the same time, new technologies would be invented to better suit the evolving users.

The inventive focus of human-machine symbiosis lies more on the interactive feedbacks between exterior technical objects and the human body than on replacing biological organs with heathier and more powerful technical organs. It is satisfied with partial automation requiring human intervention rather than progressing toward full automation. Looking at the inventions of the mouse and of the graphical user interface from Engelbart's research, it is true that both devices can be regarded as technical organs extending the human body, but the purpose of the body extensions is to serve as liaisons between the human user and the personal computer, facilitating their interactions. The body, the mind, and the senses would presumably evolve and become augmented over their interactions with and feedbacks from the personal computer. It is assumed that the body and the mind are uniquely distinct from computing devices and are augmented in different magnitudes from those of the devices.

In comparison, the cyborg imaginary focuses on the replacement of human parts with artificial limbs or artificial organs. Its end goal appears to be the replacement of the entire human with AI automation. It progresses from partial automation, which require human intervention, to full AI automation. If computing technologies can outperform the biological functions of a human body, some of which are bound to deteriorate over the

[162] This view is contrary to Donna Haraway's claim that everyone sitting in front of a computer screen is a cyborg: "By the late twentieth century, our time, a mythic time, we are all chimeras, theorized and fabricated hybrids of machine and organism; in short, we are cyborgs" (Haraway, 1991, p. 150).

[163] "Engelbart's approach is an instance of this broad conception of feedback applied to 'whole policies of behavior.' It is the reflexive application of the notion of feedback to research management as an instance of learning. In Engelbart's framework, the tool system and the human system are equally important, and the technological development of computing is associated with the human capacity to change in order to take advantage of computing as a tool" (Bardini, 2000, p. 25).

course of a lifetime, then the human body would increasingly be composed of technical components. We can find the visualization of this idea in science fiction movies such as the Star War series, in which many cyborgs are left with only the human brain and perhaps the human heart. The upgrading of cyborgs would expose the redundancy of the human body. When there is nothing left in the human that are irreplaceable by technical components, the human species would effectively become extinct. This relationship between cyborg and technological singularity is illustrated in Fabio Comparelli's computer generated simulation on human evolution (2022), in which the human species first transitions into cyborgs and then fully into AI machines.

This cyborg imaginary, interpreted with a techno-posthumanist[164] lens, is imbricated with the "survival of the fittest" mandate. Thriving for the survival of the fittest is associated with a will to technologically amplify human capabilities to conquer one obstacle after another. This will to power, from technologically conquering the exterior milieu, paradoxically leads to the reduction or elimination of human faculties, fulfilling Heidegger's abstract philosophy on the technological will to will (see Section 2.4). Those human faculties that are slower or weaker than computers or robots will eventually be supplanted. Whereas the ideal of human-computer symbiosis aims at augmenting human faculties and senses with the assistance of computing technologies, the cyborg imaginary tends to reduce or eliminate the human body. Therefore, the cyborg imaginary represents a stepping stone toward technological singularity in the techno-posthumanism.

In contrast, human-computer symbiosis does not align with techno-posthumanism. Since the goal of human-computer symbiosis is augmenting the human rather than reducing it, the vision adopts an evolutionary perspective that tends to be creative rather than eliminative. The concept of symbiosis comes from biology, meaning "the living together in more or less intimate association or close union of two dissimilar organisms" (*Symbiosis Definition & Meaning - Merriam-Webster*, n.d.). Unlike the mutual adaptation in homeostasis with equilibrium as the end goal, symbiosis is historically contingent, as the partnering organisms may change or grow in mutual dependency. The coupling is greater than the individual, like a whole is greater than its parts, but unlike a

---

[164] The techno-posthumanism here refers to Michael E. Zimmerman's adoption of the term from Ray Kurzweil's discussion on technological singularity, which Zimmerman associates with Heidegger's technological will to will. For more detailed discussion, see Section 2.4.

part in a whole, the individual in a symbiotic relationship nevertheless remains an individual. The human with its body would remain intact as an individual without having its parts replaced by technical organs. Meanwhile, it can still change and evolve over its interaction with its technological milieu.

This vision of human-computer symbiosis has been the main driver behind the technological innovations associated with personal computing and the world wide web. These innovations provide new means for doing things that free the human from unnecessary drudgery, allowing her or him to develop other strengths. For instance, to manage personal finance, spreadsheet software allows us to fully devote to the mental operations of accounting without keeping track of every menial detail in calculations. Or in the practice of writing, document software facilitates the shortening of an author's cycles between writing and editing. The habits of our mind have adapted to these tools, which open a new horizon for thinking and acting in symbiotic relations with computing devices.

9.4. The Computer Scientists' Vision of Human-AI Symbiosis

Whereas Engelbart's research on augmenting the human intellect draws from the principles of biological symbiosis, the holy grail for AI research is typically artificial general intelligence (AGI), which refers to the vision of automated machines with human-like intelligence that can operate completely independent of humans. Symbiosis and AI automation thus appear to represent two competing visions for computing research.[165] Nevertheless, AI technologies can also be deployed as symbiotic partners for the human. Many computer scientists,[166] including Kai-Fu Lee in *AI 2041*, endorse the vision of human-AI symbiosis more so than the vision of AGI, probably because AGI comes with the existential risk associated with technological singularity. As Lee remarks, "What's important is that we develop useful applications suitable for AI and seek to find human-AI symbiosis, rather than obsess about whether or when deep-learning AI will become AGI (K.-F. Lee & Chen, 2021, p. 159). This conviction lends support to optimism on a future of humanity accompanied by advances in AI: "if we have faith that human-AI

[165] See Bardini's quote (2000, p. 19) in the beginning of Section 9.3.

[166] See footnote 152.

symbiosis is much greater than the sum of the two parts, then we will work to mold AI into a perfect complement to help us 'boldly go where no one has gone before'" (K.-F. Lee & Chen, 2021, p. 537).

If AI indeed develops along the vision of symbiosis instead of AGI, what are some possible ways for humans to interact with AI? What will our lifeworld be like if our lives are predominantly in symbiotic relation with AI? The short stories in *AI 2041* (K.-F. Lee & Chen, 2021) help us imagine such a lifeworld in the year 2041, given the authors' prognosis of AI inventions in the coming years. Their prognosis is based on the understanding that AI can "detect incredibly subtle patterns within large quantities of data" (2021, p. 430). As explained in Chapter 5, this understanding is not just an expert opinion but has its basis in Solomonoff's formal proof of algorithmic probability. Hence the science fictions in *AI 2041* portray a rather realistic projection of future AI technologies. At the same time, these stories allow readers to conjure up an imaginary lifeworld to experience emotions and to reflect on human-computer symbiosis beyond categorical analysis.[167] In these projections into the future, human senses and capabilities are augmented by technological devices. For instance, in most of the stories, many characters wore XR (Extended Reality) glasses or contact lenses to gain a perception of the world beyond their natural visions. In "The Holy Driver," teenage gamers gifted with top-notch dexterity in virtual car racing were recruited to save thousands of lives remotely by continually indulging in virtual games. Their jobs were to play games in which they would drive into dangerous situations to save people, and the gamers were unaware that these virtual games were real-time simulation of the actual reality in some far-away places.

There are also passages in *AI 2041* on how humans can co-create with technology beyond its original design scheme. In "The Twin Sparrow," the protagonist, an autistic Korean kid called Silver Sparrow, co-created an immensely beautiful video artwork with his AI companion. The programmer of this AI companion, Seon, found this co-creation stunningly impressive:

> A frantic translucent video stream flooded Seon's field of vision. Different resolutions, formats, fragmented sources all edited together in a complex rhythm of time and space. Images intertwined, occluded, overwhelmed her

[167] Categorical analysis such as Kai-Fu Lee's analysis of human's relative strengths in creativity, empathy, and dexterity (See discussion in Chapter 1).

> with their visual vortex. It took a moment before Seon could distinguish anything amid the stream. Then she began to make out a few of the images: mountains, rivers, lakes, clouds, nebulae, plant veins magnified at powers of ten, irises, microstructures of chemical compounds, wind tunnel experiments captured in high-speed photography, clips from Star Trek movies, and even the day-to-day life of Fountainhead Academy. Most of the clips, however, were completely abstract or unfamiliar. There was no way Seon could begin to describe all she saw.
>
> On a hunch, Seon raised the volume on her earbuds. She heard a soft white noise like a trickle of flowing water, subtly varying with the rhythm of the visual stream. She squinted through the video layer to focus on Silver Sparrow, across from her. She understood the sound then. He could open and close his eyes, but not his ears. For children like Silver Sparrow, sensory overstimulation could become unbearable.
>
> "You made these all by yourself? They're amazing."
>
> Silver Sparrow's lip fluttered a few times, then the audio signal amplified in Seon's ear.
>
> "It was Solaris."
>
> Seon was speechless. These AI-enabled children were beyond her understanding. (K.-F. Lee & Chen, 2021, pp. 111–112)

In this passage, Silver Sparrow and Solaris, his AI companion, co-created beyond the original design of an AI companion programmed by Seon. The introduction of an AI companion into Silver Sparrow's life opened an aesthetic spacetime for Silver Sparrow to creatively maneuver and produced the animated artwork. Such aesthetic openness is brought forth by the encounter between humans and new technologies.[168]

Nevertheless, the stories in *AI 2041* implicitly reveal the problems with a research agenda purely based on human-AI symbiosis without taking into account the critiques on political economy. They paint a picture of human-AI symbiosis in a world with a predominantly capitalist ethos. In this world, people may engage with AI symbiotically even though the technologies remain alienated from the people. Our human lives would remain dependent on the good will of the elites in control of the technology. Individuals would have to struggle to maintain a healthy balance between freedom from and dependence on their hegemonic socioeconomic systems. This struggle is the main theme of "The Golden Elephant," where the protagonist lives in a world where even love

[168] This co-evolution in human and technology can be viewed as an example of transindividuation in Simondon's philosophy.

is ordained, not by parents, but by the insurance system empowered by big data and deep learning. In this imaginary world, humans and AI will co-evolve, transindividuation will still take place, and life will go on. But this is also a world in which individuals increasingly lose agency over their own lives to the elites, to the algorithmic "gods" that presumably take care of their lives.

This theme of "benevolent god" is also implicit in "The Twin Sparrow." Toward the end of the story, it was revealed that the programmer Seon has inserted a hidden piece of code into the AI companion of Silver Sparrow and into that of his twin brother Golden Sparrow, generating telepathic dream-like visions that eventually brought them back together. This element reveals how powerful is the agent who has control over the program. It also sneaks in a theme of benevolent dictatorship. The hidden programs in the AI companions led to a happy re-encounter between the separated twins, but this was only possible due to the benevolence of Seon, the AI programmer. What if Seon was wicked rather than benevolent? As a matter of fact, why should anyone accept their lives to be intervened by external agencies that influence and make personal choices for them? Should we allow our lives to be dependent on the good will of those who have more power over our lives than ourselves? As such, we ought not be satisfied with the world of human-AI symbiosis depicted in the stories of *AI 2041*.

9.5. A Philosophical Critique of Human-AI Symbiosis

How can the philosophical critiques of cybernetics in preceding chapters enlighten our perspective on a future society characterized by symbiotic relationship between the human and AI? From a philosophical perspective, the concept of symbiosis is not yet a fully developed idea and there is a need to evaluate symbiosis via other means. This section contends that the computer scientist vision of human-computer symbiosis lacks the social dimension and hence the revolutionary potential implied in Simondon's theory of individuation and in Stiegler's elaboration of Simondon's theory.

As explained in Chapter 3, cybernetics for Heidegger is destined to bring about the materialization of the essence of technology, and human-AI symbiosis is a thoroughly cybernetic idea. Human-AI symbiosis, which sees the human engaging in an open co-evolution with AI, would be regarded as a deceptive discourse that hides the further reduction of the human to standing reserves for the incessant empowerment of

AI. But as I have argued in Chapter 3, the universality of cybernetics is a mistaken characterization of the heterogeneous intellectual movement, and the conflation between cybernetics and computing technologies would efface the nuances of their relation. In more precise terms, universality is a proper characterization of the modern computers that model after the universal Turing machine, whose applicability to almost every facet of human lives is evident over the past half-century. This universality makes possible the adoption of selected cybernetic ideas in the history of computing research and development, which has bought about an epochal change of sociotechnical milieu that transcends the confinement of previous regimes. Personal computing is an appropriation of the computing technology originally targeted for mass and centralized calculation; cyberspace is an appropriation of the military-invented Internet for creative grassroot communications. While the mechanization of the living in cybernetics tends to reduce the human to its standing reserves, cybernetics is simultaneously associated with a complexity and openness that tends to liberate the human from technological alienation. The tension between reduction and openness is manifested in the dialectic between the totalitarian vision and the liberative potentiality of computers. And we have seen the pendulum swinging back and forth, from the HAL of *2001: A Space Odyssey* (Kubrick, 1968) to the 1984 Apple's MacIntosh commercial (2012), from the potential of democratization in cyberspace to surveillance capitalism and digital totalitarianism. Such a dialectic perspective is also proper for our philosophical critique of human-AI symbiosis. Advances in AI may exacerbate the penetration of surveillance capitalism and digital totalitarianism, but they also hold the potentiality of breaking loose from technological alienation.

I contend that this revolutionary potential in technical politics is lacking in the discourse of human-AI symbiosis. This lack is evident in the stories of *AI 2041*. As Fredric Jamieson mused, "it is easier to imagine the end of the world than to imagine the end of capitalism" (2003). For mainstream authors like Qiufan Chen and Kai-Fu Lee, a substantial change in the politico-economic system remains at a superficial level (e.g., a jobless environment), but they could not anticipate how AI technology may be shaped socially if it is democratized. Imagine, had stories like those in *AI 2041* been written in the 1970s, the futuristic stories might have come close to predicting the kind of software applications in businesses and governmental institutions. But they could not have foreseen the emergence of computer hackers, the popularization of personal computing, or the democratization of the Internet. Any attempts at predicting concrete future applications would inevitably appear naïve in retrospect,[169] because social imagination from the top tends to lack the transformative potentiality in the lifeworld of those from below.

In this regard, we can differentiate the computer scientists' vision of human-AI symbiosis from Simondon's proposal that humans ought to relate to machines as peers rather than as masters and slaves. Like many critics of AI today, Simondon's contemporaries were concerned with the deprivation of humanity due to industrial automation, exacerbated by the emerging computing technology. Unlike these technology critics, Simondon discerned a positive development of human-machine relation. He conceptualized the technical individual as the role for coordinating ready-to-hand tools or technical elements. In his argument, this role does not inherently belong to the human (1958/2016, pp. 77–78). With this role shifting from the human to the technology, the human is now liberated from the previous scheme where the human and the technology have been in a master-slave relationship. A new relationship is emerging where the human and the technology can relate to each other as peers (1958/2016, pp. 15–16). In the first volume of *Technics and Time* (1998), Stiegler further develops our understanding of this relationship when he constructs a technical history of what he calls "epiphylogenesis," drawing from the works of paleontologist Leroi-Gourhan along with Simondon. In this philosophical articulation of human-technology symbiosis (Beardsworth, 2010), Stiegler argues that hominization is a technical process of evolution and psychic and collective individuation, and the technical history of epiphylogenesis convincingly conveys the indispensable role of technology of hominization. Hence technology, particularly tertiary retentional system, plays a constitutive role in shaping the very definition of what is considered as human.

On the surface, human-computer symbiosis is commeasurable with the peer relationship between man and machine that Simondon proposes. Many computing devices today seem to fulfill Simondon's vision of an open calculating machine whose indeterminacy allows it to be programmable (1958/2016, pp. 31–32, 149–157). But to

[169] For example, in B. V. Bowden's *Faster than Thought* (1953), the limitations are limited to cryptography, science research, government statistics, banks/commerce, and games. These applications were largely accurate predictions of applications on the mainframe computers, but not the many revolutionary technologies that came after.

appreciate Simondon, we need to grasp the philosophy in its entirety beyond his technological examples. His philosophy of pre-individual reality, individuation and transduction, points toward a future that transcends the present. Instead of what you see is what you get, the potentiality in the transindividual relation of psychic and collective individuation goes beyond the present constraints. It points toward a historical unfolding of sociotechnical changes that transcend the existing circumstances. The human and the computing technology share a reality that is greater than whatever they may be at the current snapshot. Simondon does not advocate only an open machine but an open horizon of techno-cultural formation. The social transformation over the invention of personal computers, and over the Web and cyberspace exemplify such transcending sociotechnical changes. This open horizon, hinted by Simondon in the last pages of *MEOT*, was elaborated by Feenberg and Stiegler in their appropriation of concretization and transindividuation (see my earlier discussion in Sections 8.3, 8.4, and 8.6). Our world, with its aesthetics and beauty, meanings, and teleology, is filled with tensed relations that hold the potentiality for transcending changes. Any attempt to rethink technology, to overcome the antagonism between technics and culture, must find a way to safeguard the transcendent potentiality. In fact, this is the approach undertaken by Stiegler in his proposal for a new critique of political economy (2010a), in which his critique of the generalized proletarianization of consumptions, which results from the tertiary retentional systems of hypercapitalism, demands a radical reform of education so that youngsters can achieve a synthesis between hyperattention characterized by a rapid oscillation among different tasks and deep attention characterized as the capturing of attention by a single object (Stiegler, 2010b, p. 73).[170]

---

[170] Richard Beardsworth in “Technology and Politics: A Response to Bernard Stiegler” raises the issue of technological determinism in Stiegler’s critique of proletarianization of hyperconsumption. He explains that the “ethical question driving political innovation has, consequently, to be worked out in terms of universally coordinated, but locally determined equilibriums between growth, sustainability and equity” (216), but Stiegler’s logic of excess in his reinterpretation of Marx “ignores the need today to make small distinctions, under the canopy of political regulation, within the world as a whole” (217). At the same time, Stiegler’s “technological re-reading of Freud that flattens out the vagaries of human affect and human conscience, preventing a nuanced, comparative account of the relation between contemporary consumerism and normative thought and behaviour” (222). In short, Stiegler’s insight on the implication of tertiary retentional system on our protention becomes problematic when it is considered as the determining factor of political economy and of human affect and human conscience.

9.6. Conclusion

This chapter presents the final analysis of AI and machine learning by deliberating the possibility of democratizing AI and the openness of human-AI symbiosis. The analysis leverages the immanent critique of machine learning (Chapter 3 to Chapter 5), tracing its cybernetic origin and articulating its technical affordance based on Solomonoff's computational theories, and the philosophy of openness in cybernetics and in Simondon's philosophy (Chapter 6 to Chapter 8). In this chapter, I begin by deliberating the totalitarian threat of AI automation, which may bring about the realization of Heidegger's technological will to will. Living in an automatic society is like living within a milieu of gigantic cybernetic organisms, but these cybernetic organisms are still scientific and technical objects that attain their meanings only in relation to the experiences of ordinary humans. They lack the creative openness of Simondon's philosophy if the humans are removed from the equation. But pure automation is not the only possible goal for AI development. Humans and AI can relate to each other symbiotically. In fact, automation and symbiosis represent two competing visions for computing research. Whereas pure automation tends to reduce the human, partially automated AI, which requires human interventions to fully function, can serve to augment the human in a symbiotic relation. This computer scientists' vision of human-AI symbiosis is endorsed by the authors of *AI 2041*. They see their vision of human-AI symbiosis as potentially beneficial to our society, contrasting the vision of artificial general intelligence that leads to dystopian fear about technological singularity or about the totalitarianism of an automatic society. But the human-AI symbiosis in *AI 2041* falls short of the human-machine relation in Simondon's philosophy. Whereas human-AI symbiosis appears to be compatible with Simondon's notion of human-machine relation, in which neither the human nor the machine is the slave in subservience to the other, Simondon sees such peer relations as the realization of the pre-individual charge shared by the human and the machine, and such realization is manifested in transinindividuation of the human, the technics, and the society. Therefore, missing from the computer scientists' vision of human-AI symbiosis is the revolutionary character in Simondon's theories of concretization and transindividuation, which are adopted by Feenberg and Stiegler in their social theories on contemporary technics.[171]

To conclude, it is historical contingent regarding whether human-AI symbiosis will be realized as the enframing of the human as the standing reserve of AI technology or as the creative actualization of sociotechnical potentiality. According to Stiegler's critique of algorithmic governmentality (see Section 8.4), the trend of theoretical knowledge supplanted by automated knowledge, made possible by deep learning, is undermining the social process of negotiating knowledge and truths among stakeholders with conflictual interests. This in turn eliminates the process of disparation as the basis for psychic and collective individuation. If Stiegler is correct, algorithmic governmentality would be another example of sociotechnical innovation stabilized by recurrent causality: as it disrupts the social process of actualizing revolutionary potentiality in technical politics, this disruption results in a milieu that allows algorithmic governmentality to subsist. Therefore, even though humans may relate symbiotically to specific AI technologies, the human society may still be perpetrated by AI technologies that promote the automation of knowledge creation. A new sociotechnical innovation is needed to break away from this loop, but whether this innovation can be brought about assertively by resistant social movements or by good fortune (as in Heidegger's eschatology of Section 2.5) is a question beyond the scope of this dissertation.

[171] Today, the level of excitement and anxiety about AI has surged since the recent introduction of ChatGPT into popular culture. At the same time, its capability is no more than the AI assistants in the short stories of *AI 2041*. If we apply the arguments in this chapter to applications like ChatGPT, they can in fact augment human capacity and develop a new approach to solving daily problems. It is more user friendly than google search because the chat is conducted over exchanges in natural language. At the same time, the full potentiality of AI, its potential to disrupt incumbent sociotechnical systems, demands imagination beyond those like ChatGPT.

Conclusion

Summary of Dissertation

This dissertation presents a philosophical critique of machine learning based on an investigation into its technical lineage. It begins by explicating Heidegger's remark in *Der Spiegel* magazine that cybernetics would take the place of philosophy. As explained in Section 2.2, philosophy as western metaphysics used to provide the ground, the horizon, for determining how beings are revealed. Based on such representational systems, *Dasein* can draw associations between beings with meanings in order to make sense of the world that it has been thrown into. This traditional role of philosophy would be superseded by the "fundamental science" of cybernetics. In this regard, philosophy as western metaphysics is coming to its completion as it has found its material form in cybernetics. From his lectures on Nietzsche's will to power, Heidegger sees the history of metaphysics as a series of epochs characterized by the rise of willful subjectivity that culminates in the technological will to will (see Sections 2.3 and 2.4). Cybernetics is the materialization of the technological will to will and sets the horizon for effecting changes in the western world. No thought can escape the total domination of technology until an epochal change of collective awareness about the extreme danger of technology (see Section 2.5). No purely human effort can bring about this epochal change. Thus "only a god can save us."

Heidegger's critique of cybernetics rides on the assumption that cybernetics is a fundamental science of all sciences. Under this assumption, all scientific theories about the living and the non-living beings can be represented as cybernetics theories, and activities across all domains of reality are universally representable as cybernetic behavior. As I have contended by reading the transactions of the Macy Cybernetics Conferences in Section 3.2, the cybernetic themes of feedbacks, homeostasis, and information can indeed be appropriated to scientific disciplines across the board, but not without contradictions and conflicts typical of interdisciplinary projects. Ideas from cybernetics research served as an impetus for cross-disciplinary dialogues rather than as an agent for harmonizing different domains of knowledge into a reduced form. In fact, far from becoming a fundamental science guiding the research of all scientific disciplines, cybernetics has been unable to produce concrete and impactful scientific knowledge or theories beyond cybernetic themes such as feedbacks and homeostatis, and is therefore regarded as a historical failure from a technical perspective (See Jean-Pierre Dupuy's description of this failure in Section 3.1 and footnote 46). Yet, I believe that the true implication of cybernetics has been indirect, as it has been influential in how the computing technology has evolved, and particularly in the technical lineage of AI and machine learning. Modern-day computers are modelled after the universal Turing machine (UTM). The Church-Turing thesis suggests that it can perform any rule-of-thumb process, and that it can perform the functionality of any conceivable machine. Its universality is manifested in how computer applications find their way into every aspect of our lives. While the most straightforward application of universal computing is for speedy and complex calculations, thus replacing the factory of human computers (see Section 3.3), it is the application of cybernetic ideas in the field of computer science that has shaped how computer science has evolved (see Section 3.4). Had there been no cybernetics movement, the history of computer science would have been significantly different. As discussed in Sections 3.4 and 3.5, artificial intelligence, personal computing, and the cyberspace are branches of computer sciences that have taken up the cybernetic idea of problematizing the boundary between humans and machines. Among these subfields, AI grew out of automata studies from cybernetics research (see Section 3.5). This subfield inherits the problematization of human-machine boundary from cybernetics but departs from its primordial goal of attaining mechanistic understanding of living organisms. Instead, AI research follows Turing's outlines for machine intelligence and explores the potential of the universal computer in simulating human intelligence.

This change of focus is particularly evident in the research on machine learning as a subfield of AI. While many are aware of the cybernetic root of AI and machine learning, Chapter 4 points out the differences between the cybernetic learning machines, such as Shannon's mechanical mouse or Ashby's homeostat, and the subfield of machine learning in AI research. The differences are exemplified by the computational algorithms of Arthur Samuel's checkers playing machine and the computational theory in Ray Solomonoff's algorithmic probability. The idea that machine can "learn" has been discussed and explored in cybernetics research. As explained in Section 4.2, the goal of Claude Shannon's maze-solving mouse and W. Ross Ashby's homeostat is to find a mechanistic model of learning for the scientific understanding and mechanistic replication of the learning in living organisms, as evident in Ashby's claim that his homeostat is a model of the brain. In contrast, the pioneering works on machine learning by Samuel and by Solomonoff are based on abstract computer algorithms that have no relation to the learning mechanism in the living (See Section 4.4)

Chapter 5 explains Solomonoff's algorithmic probability as a proof that exploits the universality of UTM in the context of machine learning. Section 5.3 explains how he came up with an abstract algorithm for a UTM to learn from a universe of digital data, and proved mathematically that, given infinite time and computing resources, this algorithm guarantees the discovery of any subtle pattern in the universe of data. Section 5.4 explains the proof in non-mathematical languages to make it understandable for non-technical readers. As argued in Section 5.6, this proof turns out to be applicable to big data and deep learning, thus lending support to the non-substantiated claim by computer scientists that deep learning can discover any subtle patterns in large amount of data. Thus Solomonoff's proof can be viewed as the machine-learning version of the Church-Turing thesis that establishes the extreme outer limit of its technical affordance.

The deep learning program that recognizes subtle data patterns is also equipped with the capability to re-generate variations of the same data patterns. Generative-AI applications, such as DeepBach, rekindles curiosities about the possibility of computational creativity. To deliberate on such questions, Chapter 6 turns to the works of Simondon, who developed his philosophy of openness from Bergson's *Creative Evolution* (1922) and from the open character of cybernetics. Section 6.2 traces the openness of cybernetics to the philosophy of Norbert Wiener, arguably the most prominent pioneer of the cybernetics movement. Wiener's autobiography reveals his repudiation of the closed system of traditional sciences. For him, the essential irregularity of the universe escapes the classical repertory of analysis, in which scientific hypotheses stipulated in formal mathematical functions. The primary aim of cybernetics is to explore and to bring understanding about disorderly systems across heterogeneous environments. It is within the confine of this philosophy of openness and complexity that Wiener came up with the notions of negative feedback and homeostasis. Simondon identifies openness as the core philosophy of cybernetics, from which he developed his philosophy in *ILNFI* and *MEOT*. The eclectic nature of these two works can only be matched by the dialogues from the transactions of the Macy Cybernetics conferences. Sections 6.3 and 6.4 describe in detail how Simondon's theories are appropriations of various key concepts in cybernetics. Simondon's theories of recurrent causality, concretization, and individuation are philosophical appropriation of the cybernetic notions of negative feedbacks, homeostasis, and phylogenesis or ontogenesis. At the same time, as explained in Section 6.5, he rejects the cybernetic mechanization of the living and its blurring of the boundary between life and machine. Life and machines are independent layers of individuations with mutual influences. Rather than conflating the human and the machine, the two are capable of co-individuating or trans-individuating. In other words, Simondon's philosophy suggests a possible trajectory of co-evolution between the human, the social, and machines.

Simondon's philosophy has been characterized by some as speculative ontology ontology (Seibt & Rodogno, 2019). But as I contend in Chapter 7, Simondon's philosophy, and in particular the concept of pre-individuality, is not based on pure speculation but models the subatomic physical world, whose behavior is only observable via the assistance of modern devices. While Aristotle's model of potentiality was life, this model is outdated in light of the advances in modern sciences, which brings about a paradigm-shift on our knowledge about the physical and biological world. Thus Simondon invents a new model of potentiality from quantum theory and solid-state physics. Chapter 7 attempts to shed light on this model by referring to *Feynman Lectures on Physics* to garner some basic understanding of quantum theory and solid-state physics. This chapter explains how the paradigm of transducing the continuous into the formation of the discontinuous in quantum theory (see Section 7.2) and solid-state physics (see Section 7.3) are foundational to his conceptualization of pre-individuality and recurrent causality. Simondon models the transductive relations between the continuous and the discontinuous as a recurrence of causality between the figure and the ground (see Section 7.4). This figure-and-ground paradigm is applicable to his analysis of technology. Technological alienation is the result of a break between the figure of technical schemas and the ground of the natural, psychic, or the social milieu. Re-establishing the broken channel can therefore subvert technological alienation.

Chapter 8 compares Feenberg's and Stiegler's appropriations of Simondon's theories of concretization and transindividuation to their critical social theories and explains why the former finds appeals in the openness of concretization whereas the latter is alarmed by the closed nature of a concretized sociotechnical system. Feenberg embraces Simondon's philosophy of openness in his technical politics, in which technology can evolve openly when the voices and values from below can be incorporated into technical design (see Section 8.3). He finds Simondon's philosophy compatible with his Marcusian critical theory and adopts the theory of concretization by substituting the natural-geographic associated milieu with the social milieu. And as discussed in Section 8.2, Simondon's concept of pre-individual charge has implications similar to Hegel's potentiality, from which Marcuse develops his two-dimensional ontology. Both Simondon's pre-individual charge and Hegel's potentiality are concerned with real possibility over formal possibility, and in both, the overcoming of conflicts and tensions in inner and outer relations brings about the actualization of potentiality. Feenberg is representative of the view that computing technology is potentially liberating, which is in dialectic tension with the view that identifies its totalitarian possibility. Like Feenberg, Stiegler appropriates Simondon's theory of concretization (see Section 8.4), substituting the natural-geographic associated milieu with the social milieu, but he draws an opposite conclusion. He identifies a closed character of technical concretization: As a system concretizes, it tends toward an increasingly closed system unless changes in the natural or human milieu compel a restructuring of the technology into a new technical lineage. While Stiegler is aware that concretization can indeed transcend existing conflicts to bring about innovation, he points out that a healthy channel between the figures and ground of pre-individual charge is a necessary condition for concretization. In the sociotechnical context, the figures are technologies and the ground is the realm of reflective psyches and social critiques. Algorithmic governmentality and 24/7 capitalism undermine what Stiegler calls the "noetic time" for intellectual and spiritual reflections. Technological alienation results from this short-circuiting of the figures and the ground. Comparing Stiegler with Feenberg, their interpretations of Simondon are in fact consistent with each other (see Section 8.6). The key difference is in their assessments on the condition of the channel between social critiques and the process of inventing technologies. Feenberg points to the past movements of social resistance and emphasizes the resulting improvement over the past half century. Stiegler examines the nature of algorithmic governmentality and projects a future in which critical thoughts are decimated. The question for us is whether this dystopian future is inevitable or whether there can be alternate futures for the co-evolution of AI and society.

In my final analysis in Chapter 9, I turn to human-computer symbiosis as a competing vision to AI automation and to the cyborg imaginary for the computer scientists since the early 1970s. While the vision of human-computer symbiosis aims at augmenting the human mind, body, and senses through cybernetic feedbacks with computing devices, AI automation and the cyborg imaginary tend to reduce, replace, and eliminate human parts and faculties (see Section 9.3). The computer scientists' vision of symbiosis, as an alternative to automation, echoes the openness of cybernetics (see Section 9.4), but this computer scientists' vision nonetheless falls short of the openness that Simondon articulates in his philosophy (see Section 9.5). In Simondon's theory of transindividuation, the machine co-evolves with the human and the society by transcending prior conflicts and stagnations, and such transcendence corresponds to the revolutionary potentiality in Feenberg's technical politics.

Contributions to the Literature

Edward Feigenbaum's sentiment that AI needs a good Dreyfus (Section 1.3), together with Heidegger's dystopian critique of cybernetics (Chapter 2), frame the methodology of this dissertation: to take seriously an immanent perspective of AI, both historically and technologically, and to conduct a critical and philosophical inquiry of the technology based on this immanent perspective. As such, the primary contribution of this dissertation comes in how it draws together technical and scientific perspectives with critical philosophy of technology to formulate a constructive criticism of AI, one that aims at keeping pace with the latest technological advances. By engaging with technical and scientific knowledge, I attempt to bring further clarity to the critiques of Simondon, Feenberg, and Stiegler and explore the implication of their thoughts on artificial intelligence and machine learning.

I develop an immanent perspective of AI by tracing the intellectual traditions of cybernetics and computing (see Chapter 3 to Chapter 4) and by exploring the implications of Solomonoff's computational theories to big data and deep learning (Chapter 5). As I argue, Solomonoff's theories can be seen as a corollary to the Church-Turing thesis. Whereas the Church-Turing thesis gives an extreme outer limit of what it is possible to compute, Solomonoff's Algorithmic Probability gives an extreme outer limit on what machine learning is capable of. Since a deep neural network is Turing complete, a deep neural network can function like a computer program in identifying and regenerating data patterns. In other words, deep learning represents a shift of focus in the field of machine learning, from statistical models of correlations to the deep neural networks that can be programmed like a Turing machine. This more concise understanding of deep learning, gained by grasping the core idea behind Solomonoff's algorithmic probability, is useful in diagnosing the weaknesses of writings that downplay AI's genuine potentials. At the same time, it helps repudiate the unfettered imagination of AI enthusiasts who anticipate the presence of creativity, emotivity, and consciousness in their thriving toward artificial general intelligence.

By tracing the technical concepts from cybernetics to computing, this dissertation places a greater emphasis on the distinction between cybernetics and universal computing than their mutual influences and interwoven strands of development. This leads to a the universality of cybernetics, which underlies Heidegger's formulation of the technological will to will. While cybernetics can be distinguished from universal computing, the boundary-blurring of cybernetics led to the destabilization of the meaning of intelligence and of learning in Turing's paper on machine intelligence, which in turn led to the birth of AI as a subfield of computer science (see Sections 3.4 and 3.5). Chapter 6 examines the perspective that cybernetics is not only characterized by a reductionism of mechanizing the human, but also an openness that makes possible a critical constructivist view of technological development in Simondon's philosophy (see Section 6.6). Chapter 7 then draws from quantum and solid-state physics to reveal a deeper appreciation of the openness of Simondon's philosophy, in which potentiality is conceptualized as the ground of pre-individual charge that gives rise to the operation of individuation. Such potentiality reflects a sense of creative evolution that goes beyond the pattern recognition and regeneration in deep-learning AI. My Simondonian critique thus attempts to reveal a particular shortcoming of the kind of creativity and emotivity simulated in generative AI.

At the same time, the tension between reductionism and openness in cybernetics are relived in the dialectic tension between two research visions for AI: the universality and closed character of a totality of AI automation on the one hand, and the complexity and openness of human-AI symbiosis on the other. This dissertation further develops Simondon's philosophy to formulate a philosophical critique of human-AI symbiosis by drawing on the works of Bernard Stiegler and Andrew Feenberg. These two philosophers formulated their critiques of technology by appropriating Simondon's theories of concretization and individuation to the social realm. I attempted to clarify how their thoughts relate to one another, which can be illuminated by recognizing the simultaneous open and closed character of technical concretization. This illumination reiterates the dialectic tension in the pharmacology of technology that may be poisonous or curative to the human and social world. It leads to the question of whether human-computer symbiosis or human-AI symbiosis can be a liberative alternative to the enframing of AI automation. On the surface, a symbiotic partnership between humans and AI appears to be compatible with the philosophy of Simondon, who holds the view that humans ought to relate to machines as peers rather than as masters and slaves. But as I argue in Section 9.5, missing in the computer scientist vision of human-AI symbiosis is the social dimension in Simondon's philosophy, as identified by Feenberg and Stiegler. Simondon intends to overcome not only the opposition between man and machine, but more importantly, the opposition between culture and technics. He advocates not only an open machine but also an open horizon of techno-cultural formation. Unlike the computer scientist vision of human-computer symbiosis, Simondon's philosophy points to a historical unfolding of sociotechnical changes that transcend the contradictions of existing circumstances.

Recommendations/Areas of Future Research

This dissertation examines a technology that has become the focus of governments' strategies across the globe with implications on multiple fronts, from military warfare to the shape of the technical milieu in our lifeworld. The corresponding funding support guarantees a fast pace of development in AI research for the coming years. Moreover, the availability of generic machine-learning model such as the pre-trained GPT model, which can be adapted to specific usages by fine-tuning the model, will bring into existence many new AI applications. One such adaptation is the widely popular ChatGPT. In the fine-tuning process of ChatGPT, additional layers are added on top of the pre-trained GPT neural network to adapt it to the specific task of generating natural language responses in a conversational setting. There are also many commercial products, including those by Google and Microsoft, for creating domain-specific chatbots. As evident by Elon Musk's indication to develop TruthGPT, 'a maximum truth-seeking AI' beyond the politically correct ChatGPT, AI chat agent is becoming the heir apparent to Internet Search as the final arbiter of "truth" (see Mehta, 2003). We can anticipate that many issues with social implications will come with the fast pace of AI development.

One such issue is the problem of AI alignment, which has become increasingly pertinent due to the ever-growing complexity and size of AI models and training datasets. An aligned AI system advances the designers' intended goals whereas a misaligned AI system is competent at advancing some objectives, but not the intended one (Russell et al., 2022, pp. 31–34). Due to the way AI models are typically trained with massive amount of data, the data scientists themselves cannot guarantee a resulting model that aligns with their intended goals. These problems affect existing commercial systems such as robots (Franklin & Ashton, 2022; Kober et al., 2013), language models (Bommasani et al., 2022; Ouyang et al., 2022), autonomous vehicles (Knox et al., 2022), and social media recommendation engines (Bommasani et al., 2022; Russell et al., 2022, pp. 31–34; Stray, 2020). Such examples are frequently found in gaming development. For instance, in the game "CoastRunners" developed by OpenAI, the goal for the AI agent is to finish the race faster than other players along a pre-specified route. To ensure the players to follow the route of the race, the game would reward players with points if they hit targets laid out along that route. It turns out that the reinforcement learning agent found a way to get higher points without finishing the race: by looping and crashing into the same targets indefinitely (Clark & Amodei, 2016). With automated applications that build on large language models like GPT-4, AI can perform online actions like an avatar represented by a human user.[172] If it becomes misaligned, it may be susceptible to power-seeking and poses the most catastrophic misalignment risks (Bommasani et al., 2022; Carlsmith, 2022; Wei et al., 2022). This problem is exacerbated by the recent democratization of generic large language models that anyone can fine-tune or build software on.[173]

My approach in this dissertation, which combines an immanent understanding of the technology and brings this understanding into conversation with critical philosophy, seems appropriate for tackling new issues like that of AI misalignment in large language models. For instance, on the deliberation about the existential risk with AI misalignment,

---

[172] See footnote 2.

[173] Hinton is particular concerned that such a move is morally irresponsible and can bring chaos into the online world (Joseph Raczynski, 2023).

we need to understand the potential capability and limitation of large language models. This would require an investigation into the machine-learning model called “transformer” (see Vaswani et al., 2017). The affordance of deep learning explained in this dissertation would contribute to the technical understanding of the transformer model, which consists of multiple deep neural networks. The knowledge of how the transformer model works is essential for a critique of the claim that large language models can understand human language and can reason as well as or better than a human being,[174] or for examining the fascination on whether putting large-language model in a self-reflexive loop would imitate the self-consciousness of human being (AI Explained, 2023; bycloud, 2023; QuestionAlt1, 2023).

When facing a new controversy about a new sociotechnical phenomenon, such as the issue of AI misalignment coupled with the democratization of generic large language models, a recommended approach is to move past the dichotomy between the technical culture and humanities by synthesizing a technical inquiry with the perspective of critical philosophical works. Such an approach can bring into conversation the AI community and the critical thinkers of continental philosophy, two communities that have been mostly isolated from the other epitimozed by the conflict between Dreyfus and the early AI pioneers. A critical perspective of technology from continental philosophy can engage in a dialogue with the immanent understanding of new technologies. From the disapproval of Heidegger’s technological will to will to the philosophy of openness in Wiener’s cybernetics and in Simondon’s works, the deliberation of continental philosophy in this dissertation identifies the positive potentials in human-AI symbiosis in which new technologies bring out latent human potentials prior to their coming into beings. This philosophical perspective is critical of a reductionist vision in which human-technology relations tends toward a master-and-slave hierarchy. It points toward an open horizon that guides humans and technologies to grow in a mutually beneficial relation.

[174] Geoffrey Hinton suggests that these language models will soon possess an IQ of 210 (Joseph Raczynski, 2023).

Coda: Suggestions of pharmacological AI technologies

Up to this point, I have refrained from proposing potential AI applications that liberate rather than alienate. Nonetheless, it would be helpful to envision what liberative AI technologies may look like, however naïve my suggestions may look in hindsight. Below are my proposals of three such AI applications. In coming up with these applications, I am reiterating the perspective of this dissertation that technological alienation is not inherent in human-AI relations. These AI applications are possible countermeasures against the problems of surveillance, algorithmic governmentality, and the technological domination of gigantic cybernetics organism. In other words, they exhibit how AI can be pharmacological (see Section 8.4), being capable of both poison and cure. Utilizing the AI's affordance of pattern recognition, they are deep-learning solutions of the problems introduced through the breakthrough of deep learning, exemplifying how technology can be transformed to address the problem of technology, exemplifying what Feenberg and Ihde call the gestalt switch (see Section 8.3).

The first AI application is a health monitoring device for mental health. While many health monitoring systems are implemented with centralized servers for surveillance, it is technically feasible to implement one without surveillance. Assuming a health monitoring system where private data are free from surveillance, it can be beneficial for patients or elderlies to keep track of their physical health (heartbeats, blood pressure, and so on). In fact, since the affordance of deep learning is pattern recognition and not tied to surveillance, this affordance can be implemented as an antidote against the social ailments that surveillance capitalism has introduced. One possible antidote is an application that tracks and raises alerts on a person's mental health. For instance, it can detect patterns of compulsive buying behavior, excessive web browsing or gaming activities, or an extreme imbalance in the choices of food intake. This would help foster a reflective perspective on a person's well-being, recovering the "noetic time"[175] for critical thoughts and reflections that are marginalized under algorithmic governmentality according to Stiegler. This application should be especially appealing for parents who are concerned about the unguarded influences of media exposure on their children.

[175] See Section 8.4 for how Stiegler defines noetic time.

The second AI application addresses the issue of opacity. Earlier in this dissertation, I brought up Stiegler's critique of algorithmic governmentality in Section 8.4 and Hui's critique of gigantic cybernetic organisms in Section 9.2. The opacity of AI is central to both critiques. The inability for a human to understand AI's decision-making, presumed to be superior to humans, makes it impossible to critique the rationality of the decision. This in turn put AI and humans in a different scale, an unfathomable gigantic cybernetic organism versus a feeble and powerless human weakling. But a peer relationship between AI and humans is a necessary condition for the human-AI symbiosis advocated in Section 9.5. Hence the opacity of AI may bring an imbalance to research on human-AI symbiosis. The problem is, the more powerful is a machine-learning model and the more complex is the model's architecture, the more opaque would be the reasoning behind the model's decision (see Lo, 2022). Any regulation to enforce the complete transparency of AI decision-making would be equivalent to the outlawing of AI.

Nonetheless, it can still be helpful to get a rough idea about the reasoning behind AI's decision-making. This would be analogous to attempts at unveiling the reasoning behind human intuitions. Athletes, musicians, and scientists are often asked to explain how they arrive at their physical, aesthetic, or mental intuitions. The same is true for autistic savants. Ever since the movie *Rain Man* (1988) came out, it has become a common knowledge that autistic savants can perform mind-boggling mathematical calculations, but no one understands why, at least until recently. Daniel Tammet, an autistic savant who can recall pi to 22,514 decimal places but cannot tell right from left, unveils the mental activities behind his extraordinary ability (Johnson, 2005). In the past, no one knows because most autistic savants cannot tell how they do what they do. But unlike other savants, Tammet can describe what he sees in his head:

> Since his epileptic fit, he has been able to see numbers as shapes, colours and textures. The number two, for instance, is a motion, and five is a clap of thunder. "When I multiply numbers together, I see two shapes. The image starts to change and evolve, and a third shape emerges. That's the answer. It's mental imagery. It's like maths without having to think" (Johnson, 2005).

A deep-learning AI is like the mind of an autistic savant. Even though it is impossible to articulate its precise means for arriving at decisions, it is possible to extract an approximate description and explanation of how they come to such decisions. Such a translation between machine reasoning and human-scale understanding may involve the training of a second AI model, feeding it with the past decisions of the first model. This post-hoc analysis can identify patterns and possibly unveil a rough sketch of how the model arrives at its decisions, in the same way that Daniel Tammet roughly sketches out how an autistic savant can perform mind-boggling mathematic calculations. Indeed, there are projects in explainable AI (XAI) research that explore how post-hoc analysis may overcome the issue of opacity in machine learning (E. Lee et al., 2021). This would in turn contribute to research on human-AI symbiosis.

Like my first proposal, my third proposed AI application is a countermeasure against surveillance capitalism and a pharmacological solution based on the pattern recognition technology in deep learning. The application can perform diagnosis on the biases in online recommendations. These biases steer users' desires and behaviors toward the goals endorsed by online platforms, effectively modelling after the goal-oriented feedback loops in cybernetics. In Stiegler's terms, the platforms manipulate users' protention and marginalize their noetic time for critical thoughts and reflections (see Section 8.4). But such biases, however complex, should correspond to some data patterns. Therefore, a third-party plugin for a web browser can record a historical trace of the recommendations for each website, and the recorded data can then be used to train an AI model to recognize the patterns of how each online platform manipulates a specific user. The analysis can help users foster an awareness of how online recommendations limit their freedom and help them get a rough idea about the online profiles concealed from them. This reverse engineering of Internet surveillance involves the translation of machine reasoning[176] to human-scale understanding that is technically similar to the XAI project of post-hoc analysis. I would imagine that people ought to be interested in gaining some understanding about their online profiles, regardless of the varying degrees of their dependence on online recommendations. A collective awareness can in turn become the impetus crucial for real actions in technical politics.

[176] Collaborative filtering is one possible algorithm for implementing a recommender system. It computes rating predictions by autonomously discovering features that characterize the content of an item, and these features may differ from typical human descriptions. For instance, a person may describe a given movie as a romantic comedy while the algorithm would characterize it with more fine-grained features. Collaborative filtering auto-discover relevant "features," the meanings of which are very difficult for humans to decipher.

These three proposed applications bring together some of the key arguments raised in this dissertation. For instance, AI is not inherently subservient to humans nor vice versa, and the affordance of deep-learning AI is not surveillance and algorithmic governmentality, which are only particular applications of the affordance of pattern recognition in a capitalist or totalitarian society. The affordance to recognize data patterns can equally be utilized as countermeasures against the capitalist and totalitarian social issues that have been exacerbated by deep-learning AI. In that sense, the social problems aggravated by technological advances are met with countermeasures made possible by the same advances. At the same time, whether the proposed applications can be the impetus for the transindividuation in Simondon's philosophy would not be known until they are deployed in actual practices and undergo a subsequent process of socially shaping by the technology users.

References

1984 Apple's Macintosh Commercial (HD). (2012, February 1). https://www.youtube.com/watch?v=VtvjbmoDx-I

A Short List of Gilbert Simondon's Vocabulary. (2007, November 28). *Fractal Ontology*. https://fractalontology.wordpress.com/2007/11/28/a-short-list-of-gilbert-simondons-vocabulary/

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*(1), 147–169.

Actuality and Potentiality in Aristotle's Philosophy | UPSC Notes. (n.d.). *IAS EXPRESS*. Retrieved October 21, 2023, from https://www.iasexpress.net/modules/1-7-actuality-and-potentiality-in-aristotles-philosophy/

Adrian German (Director). (2018, June 26). *Ray J Solomonoff Midwest NKS 2005 Lecture*. https://www.youtube.com/watch?v=CfjkqpzDkCs

Akrich, M. (2010). The De-Scription of Technical Objects. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (Nachdr.). MIT Press.

Allen, E. (n.d.). *No evidence of long-term welfare problems with electronic containment of cats*. Retrieved June 3, 2024, from https://phys.org/news/2016-09-evidence-long-term-welfare-problems-electronic.html

Alter, M. P., & Caputo, J. D. (1976). Only a God can save Us: Der Spiegel's Interview with Martin Heidegger. *Philosophy Today*, *20*(4), 267–284. https://doi.org/10.5840/philtoday197620422

Anderson, C. (2008, June 23). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Appadurai, A. (Ed.). (2013). *The Future as cultural fact: Essays on the global condition*. New York : Verso Books.

As AutoGPT released, should we be worried about AI? (2023, May 10). https://cosmosmagazine.com/technology/autogpt-ai-bmj-existential-threat/

Ashby, M. A., on behalf of the Estate of W. Ross. (1948). *W. Ross Ashby's 1948 Homeostat*. Own work. https://commons.wikimedia.org/wiki/File:W._Ross_Ashby%27s_1948_Homeostat.jpg

Ashby, W. R. (1948). Design for a Brain. *Electronic Engineering*, *20*, 379–383.

Ashby, W. R. (1960). *Design for a brain; the origin of adaptive behavior*. New York, Wiley. http://archive.org/details/designforbrainor00ashb (Original work published 1952)

Ashby, W. R. (2003a). Homeostasis. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 593–619). Diaphanes. (Original work published 1952)

Ashby, W. R. (2003b). Mechanical Chess Player. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 651–653). Diaphanes. (Original work published 1952)

Ashby, W. R., Shannon, C. E., & McCarthy, J. (Eds.). (1972). *Automata studies* (5th printing). Princeton University Press. (Original work published 1956)

Asimov, I. (2008). *I, Robot*. Bantam Spectra. (Original work published 1950)

Aspray, W. (1990). *Computing before computers*. Iowa state university press.

Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 26–33.

Bardin, A., & Ferrari, M. (2022). Governing progress: From cybernetic homeostasis to Simondon's politics of metastability. *The Sociological Review*, *70*(2), 248–263.

Bardini, T. (2000). *Bootstrapping: Douglas Engelbart, coevolution, and the origins of personal computing*. Stanford University Press.

Bateson, G. (1987). From Versailles to Cybernetics. In *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology* (pp. 475–483). Aronson.

Bateson, G. (2003). The Position of Humor in Human Communication. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 541–574). Diaphanes. (Original work published 1952)

Beardsworth, R. (2010). Technology and politics: A response to Bernard Stiegler. *Cultural Politics*, *6*(2), 181–199.

Becks, E., & Weis, T. (2022). Nudging to Improve Human-AI Symbiosis. *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops)*, 132–133.

Bell Lab. (n.d.). *Claude Shannon's Electromechanical Mouse*. Retrieved May 18, 2024, from https://www.bell-labs.com/claude-shannon/

Bell Labs Advances Intelligent Networks. (2012, July 22). https://web.archive.org/web/20120722011711/http://www.alcatel-lucent.com/wps/portal/%21ut/p/kcxml/04_Sj9SPykssy0xPLMnMz0vM0Y_QjzKLd4w39w3RL8h2VAQAGOJBYA%21%21?LMSG_CABINET=Bell_Labs&LMSG_CONTENT_FILE=News_Features%2FNews_Feature_Detail_000025

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Berghoff, C., Biggio, B., Brummel, E., Danos, V., Doms, T., Ehrich, H., Gantevoort, T., Hammer, B., Iden, J., Jacob, S., Khlaaf, H., Komrowski, L., Metzen, J. H., Neu, M., Petsch, F., Poretschkin, M., Samek, W., Schäbe, H., von Twickel, A., … Wiegand, T. (2021). *Whitepaper: Towards Auditable AI Systems*. 32.

Bergson, H. (1922). *Creative Evolution* (A. Mitchell, Trans.). Macmillan and Co. Limited.

Bhaimiya, S. (2023, February 1). *ChatGPT could hypothetically get hired as an entry level coder if it interviewed at Google, internal document reportedly says*. Business Insider. https://www.businessinsider.com/chatgpt-google-could-get-hired-as-coder-internal-document-report-2023-2

Bijker, W. E., Hughes, T. P., & Pinch, T. (Eds.). (1987). *The Social construction of technological systems: New directions in the sociology and history of technology*. MIT Press.

Birch, H. G. (2003). Communication Between Animals. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 446–473). Diaphanes. (Original work published 1951)

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. http://arxiv.org/abs/2108.07258

Borgmann, A. (1988). *Technology and the character of contemporary life: A philosophical inquiry* (Paperback ed., 3. pr). Univ. of Chicago Pr.

Bowden, B. V. (Ed.). (1953). *Faster than Thought: A Symposium on Digital Computing Machines*. Sir Issac Pitman & Sons, Ltd. https://archive.org/details/faster-than-thought-b.-v.-bowden/

Bowker, G. C. (2008). *Memory practices in the sciences* (1. paperback ed). MIT.

Brand, S. (2004). For God's Sake, Margaret. Conversation with Gregory Bateson and Margaret Mead. In C. Pias (Ed.), *Cybernetics/Kybernetik—The Macy Conferences 1946-1953: Essays and documents: Vol. II* (1. Aufl, pp. 301–312). Diaphanes. (Original work published 1976)

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, *8*(1), 2053951720983865. https://doi.org/10.1177/2053951720983865

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Brownlee, J. (2019, March 3). What is a Hypothesis in Machine Learning? *MachineLearningMastery.Com*. https://machinelearningmastery.com/what-is-a-hypothesis-in-machine-learning/

Campbell, D. (2013). The Semimeasure Property of Algorithmic Probability–"Feature" or "Bug"? In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence* (pp. 79–90). Springer.

Campbell-Kelly, M., & Aspray, W. (2004). *Computer: A history of the information machine* (10. Dr.). BasicBooks.

Čapek, K. (1920). *RUR-Rossum's Universal Robots: Rossumovi univerzln roboti*. Aventinum.

Carabantes, M. (2020). Black-box artificial intelligence: An epistemological and critical analysis. *AI & SOCIETY*, *35*(2), 309–317.

Carlsmith, J. (2022). *Is Power-Seeking AI an Existential Risk?* (arXiv:2206.13353). arXiv. https://doi.org/10.48550/arXiv.2206.13353

Cellan-Jones, R. (2014, December 2). Hawking: AI could end human race. *BBC News*. https://www.bbc.com/news/technology-30290540

Chabot, P. (2013). *The philosophy of Simondon: Between technology and individuation* (A. Krefetz & G. Kirkpatrick, Trans.). Bloomsbury Academic.

Chan, L. (2021). Explainable AI as Epistemic Representation. *Overcoming Opacity in Machine Learning*, 7.

Chow, A. R., & Perrigo, B. (2023, February 16). *The AI Arms Race Is On. Start Worrying*. Time. https://time.com/6255952/ai-impact-chatgpt-microsoft-google/

Chun, W. H. K. (2013). *Programmed visions: Software and memory* (1. MIT Press paperback ed). MIT Press.

Church, A. (1936). An Unsolvable Problem of Elementary Number Theory. *American Journal of Mathematics*, *58*(2), 345. https://doi.org/10.2307/2371045

Clark, J., & Amodei, D. (2016, December 21). *Faulty reward functions in the wild*. https://openai.com/research/faulty-reward-functions

Clynes, M. E., & Kline, N. S. (1960). Cyborgs and Space. Astronautics. September 1960. In *The cyborg handbook* (pp. 29–33). Routledge New York.

Combes, M. (2013). *Gilbert Simondon and the philosophy of the transindividual*. MIT Press.

Comparelli, F. (2022, October 2). *Digital content creator on Instagram: "where are we going?"* Instagram. https://www.instagram.com/reel/CjN9uibj_MK/

Cordeschi, R. (2011). *The discovery of the artificial: Behavior, mind and machines before and beyond cybernetics*. Springer.

Cutland, N. (1980). *Computability, an introduction to recursive function theory*. Cambridge University Press.

Dartmouth workshop. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Dartmouth_workshop&oldid=1103155760

Davis, B. W. (2007). *Heidegger and the will: On the way to Gelassenheit*. Northwestern University Press.

de La Bellacasa, M. P. (2017). *Matters of care: Speculative ethics in more than human worlds* (Vol. 41). U of Minnesota Press.

Dick, P. K. (1972). *Do androids dream of electric sheep?* Panther. (Original work published 1968)

Dlott, D. D. (2003). Fast molecular processes in energetic materials. In *Theoretical and computational chemistry* (Vol. 13, pp. 125–191). Elsevier.

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books, a member of the Perseus Books Group.

Dowe, D. L. (2013). Introduction to Ray Solomonoff 85th memorial conference. In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence* (pp. 1–36). Springer.

Dreyfus, H. L. (1965). Alchemy and Artificial Intelligence. *Santa Monica, CA: Rand Corporation*.

Dreyfus, H. L. (1972). *What computers can't do; a critique of artificial reason* (1st ed.). Harper & Row.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT Press.

Dunne, A., & Raby, F. (2013). *Speculative everything: Design, fiction, and social dreaming*. MIT press.

Dupuy, J.-P. (2000). *The mechanization of the mind: On the origins of cognitive science*. Princeton University Press.

Durrant, M. (2015). *Aristotle's De anima in focus*. Routledge.

Edwards, P. N. (1997). *The closed world: Computers and the politics of discourse in Cold War America*. MIT press.

Ekbia, H. R. (2008). *Artificial dreams: The quest for non-biological intelligence*. Cambridge University Press Cambridge.

Elias, J. (2023a, April 17). *Google CEO Sundar Pichai warns society to brace for impact of A.I. acceleration, says 'it's not for a company to decide'*. CNBC. https://www.cnbc.com/2023/04/17/google-ceo-sundar-pichai-warns-society-to-brace-for-impact-of-ai-acceleration.html

Elias, J. (2023b, May 1). *"Godfather of A.I." leaves Google after a decade to warn society of technology he's touted*. CNBC. https://www.cnbc.com/2023/05/01/godfather-of-ai-leaves-google-after-a-decade-to-warn-of-dangers.html

Ethics | Definition, History, Examples, Types, Philosophy, & Facts | Britannica. (2023, May 2). https://www.britannica.com/topic/ethics-philosophy

Fainman, A. A. (2019). The problem with Opaque AI. *The Thinker*, *82*(4), 44–55.

Feenberg, A. (1995). *Alternative modernity: The technical turn in philosophy and social theory*. University of California Press.

Feenberg, A. (1999). *Questioning technology*. Routledge.

Feenberg, A. (2002). *Transforming technology: A critical theory revisited*. Oxford University Press.

Feenberg, A. (2005). *Heidegger and Marcuse: The catastrophe and redemption of history*. Routledge.

Feenberg, A. (2014). *The philosophy of praxis: Marx, Lukács, and the Frankfurt School* (New edition). Verso.

Feenberg, A. (2017a). *Technosystem: The social life of reason*. Harvard University Press.

Feenberg, A. (2017b). *Technosystem: The social life of reason*. Harvard University Press.

Feenberg, A. (2023). *The ruthless critique of everything existing: Nature and revolution in Marcuse's philosophy of Praxis*. Verso.

Feynman, R. P., Leighton, R. B., & Sands, M. L. (2011). *The Feynman lectures on physics* (New millennium ed). Basic Books.

Forest, K. Z. (2016). *Whats your dream pet?* https://www.playbuzz.com/katezebraforest10/whats-your-dream-pet

Franklin, M., & Ashton, H. (2022). Preference Change in Persuasive Robotics. *arXiv Preprint arXiv:2206.10300*.

Future of Life Institute. (2017). *Asilomar AI principles*. Future of Life Institute. https://futureoflife.org/open-letter/ai-principles/

Galloway, A. R. (2006). *Protocol: How control exists after decentralization*. MIT Press.

Garelli, J. (2020). In T. Adkins (Trans.), *Individuation in Light of Notions of Form and Information* (pp. xv–xxviii). Univ Of Minnesota Press. (Original work published 1964)

Geoghegan, B. D. (2023). *Code: From information theory to French theory*. Duke University Press.

Gerard, R. W. (2003). Some of the Problems Concerning Digital Notions in the Central Nervous System. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 171–202). Diaphanes. (Original work published 1950)

Gibson, W. (2010). Neuromancer (1984). In *Crime and Media*. Routledge.

Gilbert Simondon. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Gilbert_Simondon&oldid=1144090297

Grace, K., & Weinstein-Raun, B. (2022). *2022 Expert Survey on Progress in AI* [Survey]. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/

Grey-Walter, W. (2003). Studies on Activity of the Brain. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 689–696). Diaphanes. (Original work published 1953)

Grier, D. A. (2005). *When computers were human*. Princeton University Press.

Harari, Y. N. (2018a). *21 Lessons for the 21st Century*. Random House.

Harari, Y. N. (2018b). Why Technology Favors Tyranny. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/

Haraway, D. (1991). Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. *Simians, Cyborgs and Women: The Reinvention of Nature*, 149–181.

Harris, T., & Raskin, A. (2023). *The AI Dilemma*. https://www.humanetech.com/podcast/the-ai-dilemma?ref=warpnews.org

Hartley, J. R. (2006). Some Psychological Aspects of Computer-Assisted Learning and Teaching. *PLET: Programmed Learning & Educational Technology*. https://doi.org/10.1080/1355800850220206

Hayles, K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.

Heaven, W. D. (2023, April 6). *ChatGPT is going to change education, not destroy it*. MIT Technology Review.

https://www.technologyreview.com/2023/04/06/1071059/chatgpt-change-not-destroy-education-openai/

Heidegger, M. (1953). The Question Concerning Technology. In *Basic writings: From Being and time (1927) to The task of thinking (1964)* (Rev. and expanded ed). HarperSanFrancisco.

Heidegger, M. (1968). *What Is Called Thinking?* (J. G. Gray & F. D. Wieck, Trans.). Perennial.

Heidegger, M. (1969). *Discourse on thinking* (J. M. Anderson & E. H. Freund, Trans.). Harper & Row.

Heidegger, M. (1973a). Metaphysics as History of Being. In J. Stambaugh (Trans.), *The end of philosophy* (1st ed., pp. 1–54). Harper & Row.

Heidegger, M. (1973b). Overcoming Metaphysics. In J. Stambaugh (Trans.), *The end of philosophy* (1st ed., pp. 84–110). Harper & Row.

Heidegger, M. (1973c). Sketches for a History of Being as Metaphysics. In J. Stambaugh (Trans.), *The end of philosophy* (1st ed., pp. 55–74). Harper & Row.

Heidegger, M. (1977a). *The question concerning technology, and other essays*. Garland Pub.

Heidegger, M. (1977b). The Word of Nietzsche: "God is Dead." In *The question concerning technology, and other essays* (pp. 53–112). Garland Pub.

Heidegger, M. (1983). *Denkerfahrungen, 1910-1976*. Vittorio Klostermann.

Heidegger, M. (1985). *Schelling's treatise on the essence of human freedom*. Ohio University Press.

Heidegger, M. (1991). *Nietzsche Vol. 1: The Will to Power as Art* (D. F. Krell, Trans.; 1st HarperCollins pbk. ed). HarperSanFrancisco.

Heidegger, M. (1992a). *Gelassenheit* (10th ed.). Neske.

Heidegger, M. (1992b). *Parmenides* (A. Schuwer & R. Rojcewicz, Trans.). Indiana university press.

Heidegger, M. (1993). *Nietzsche Vol. 3: The will to power as knowledge and as metaphysics* (D. F. Krell, J. Stambaugh, & F. A. Capuzzi, Trans.; Repr., 3. pr). Harper.

Heidegger, M. (2002). The End of Philosophy and the Task of Thinking. In J. Stambaugh (Trans.), *On Time and Being* (pp. 55–73). The University of Chicago.

Heidegger, M. (2007). *Gesamtausgabe. 77: Abt. 3, Unveröffentlichte Abhandlungen, Vorträge - Gedachtes: Feldweg-Gespräche (1944/45): [erdachte Gespräche 1944/45] / [hrsg. von Ingrid Schüßler]* (I. Schüßler, Ed.; 2., durchges. Aufl). Klostermann.

Heidegger, M. (2010). *Being and time* (J. Stambaugh & D. J. Schmidt, Trans.). State University of New York Press. (Original work published 1927)

Hertz, J., Krogh, A., & Palmer, R. G. (1999). *Introduction to the theory of neural computation*. Addison-Wesley.

History of artificial intelligence—Wikipedia. (n.d.). Retrieved October 3, 2022, from https://en.wikipedia.org/wiki/History_of_artificial_intelligence#Mythical,_fictional,_and_speculative_precursors

Hofstadter, D. R. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.* Basic books.

Hui, Y. (2011, September 19). The notion of information in Simondon [1]. *Digital Milieu*. https://digitalmilieu.net/119/the-notion-of-information-in-simondon-1/

Hui, Y. (2014, June 26). *Yuk Hui: Commentary on Thierry Bardini and Anne Fagot-Largeault's Conversation*. Theory, Culture & Society | Global Public Life. https://www.theoryculturesociety.org/blog/responses-yuk-hui-commentary-on-thierry-bardini-and-anne-fagot-largeaults-conversation

Hui, Y. (2019). *Recursivity and contingency*. Rowman & Littlefield International.

Hui, Y. (Ed.). (2021). *Philosophy After Automation*. *65*(2).

Huxley, A. (1932). Brave new world. 1932. *London: Vintage*.

Hwang, D. (2023, March 21). *ChatGPT's first book in Korea explores '45 Ways to Find a Purpose in Life.'* https://www.koreaherald.com/view.php?ud=20230321000711

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.

Iliadis, A. (2013). Informational ontology: The meaning of gilbert Simondon's Concept of individuation. *Communication+ 1*, *2*(1), 1–19.

Iliadis, A. (2015). Two examples of concretization. *Platform: Journal of Media and Communication*, *6*, 86–95.

Influenza in Cats | CDC. (2024). https://www.cdc.gov/flu/other/flu-in-cats/index.html

Jameson, F. (2003). Future City. *New Left Review*, *21*, 65–79.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Johnson, R. (2005, February 12). A genius explains. *The Guardian*. https://www.theguardian.com/theguardian/2005/feb/12/weekend7.weekend2

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Joseph Raczynski (Director). (2023, May 4). *Possible End of Humanity from AI? Geoffrey Hinton at MIT Technology Review's EmTech Digital*. https://www.youtube.com/watch?v=sitHS6UDMJc

Joy, B. (2000, April 1). Why the Future Doesn’t Need Us. *Wired*. https://www.wired.com/2000/04/joy-2/

Kan, M. (2023, March 15). *GPT-4 Was Able To Hire and Deceive A Human Worker Into Completing a Task*. PCMAG. https://www.pcmag.com/news/gpt-4-was-able-to-hire-and-deceive-a-human-worker-into-completing-a-task

Knox, W. B., Allievi, A., Banzhaf, H., Schmitt, F., & Stone, P. (2022). *Reward (Mis)design for Autonomous Driving* (arXiv:2104.13906). arXiv. http://arxiv.org/abs/2104.13906

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, *32*(11), 1238–1274. https://doi.org/10.1177/0278364913495721

Koch, C. (2016, March 19). *How the Computer Beat the Go Master*. Scientific American. https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/

Kubie, L. S. (2003). Communication Between Sane and Insane: Hypnosis. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 416–445). Diaphanes. (Original work published 1951)

Kubie, L. S. (2003). The Place of Emotions in the Feedback Concept. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 575–592). Diaphanes. (Original work published 1952)

Kubie, L. S. (2003). The Relationship of Symbolic Function in Language Formation and in Neurosis. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 307–325). Diaphanes. (Original work published 1950)

Kubrick, S. (Director). (1968). *2001: A Space Odyssey*. Metro-Goldwyn-Mayer Studio Inc.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed). University of Chicago Press. (Original work published 1962)

Kurzweil, R. (2000). *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* (1 edition). Penguin Books.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Viking.

Lee, E., Taylor, H., Hiley, L., & Tomsett, R. (2021). Technical Barriers to the Adoption of Post-hoc Explanation Methods for Black Box AI models. *Overcoming Opacity in Machine Learning*, 12.

Lee, K.-F., & Chen, Q. (2021). *AI 2041* (First edition). Currency.

LeVine, H. (2010). *The great explainer: The story of Richard Feynman* (1st ed). Morgan Reynolds Pub.

Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications* (Vol. 3). Springer.

Licklider, J. C. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *1*, 4–11.

Lindberg, S. (2019). Being with technique–technique as being-with: The technological communities of Gilbert Simondon. *Continental Philosophy Review*, *52*(3), 299–310.

Lo, F. T. H. (2016). *The Dialectic of Open Technology* [Thesis, Communication, Art & Technology: School of Communication]. http://summit.sfu.ca/item/16386

Lo, F. T. H. (2019). The Dilemma of Openness in Social Robots. *Techné: Research in Philosophy and Technology*, *23*(3), 342–365. https://doi.org/10.5840/techne20191126107

Lo, F. T. H. (2022). The paradoxical transparency of opaque machine learning. *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01616-7

Longo, A. (2022). *Le jeu de l'induction: Automatisation de la connaissance et réflexion philosophique*. Éditions Mimésis.

Lords, H. O. (2018). AI in the UK: Ready, willing and able? *Retrieved August*, *13*, 2021.

Lovink, G. (2019, September). *Cybernetics for the Twenty-First Century: An Interview with Philosopher Yuk Hui - Journal #102 September 2019 - e-flux*. https://www.e-flux.com/journal/102/282271/cybernetics-for-the-twenty-first-century-an-interview-with-philosopher-yuk-hui/

Lukács, G. (2013). *History and class consciousness: Studies in Marxist dialects* (R. Livingstone, Trans.). MIT Press.

Lupton, D., & Watson, A. (2022). Creations for Speculating About Digitized Automation: Bringing Creative Writing Prompts and Vital Materialism into the Sociology of Futures. *Qualitative Inquiry*, *28*(7), 754–766.

MacKay, D. M. (2003). In Search of Basic Symbols. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 480–510). Diaphanes. (Original work published 1951)

MacKenzie, D. A. (2004). *Mechanizing proof: Computing, risk, and trust*. MIT.

Mahmud, B. U., Hong, G. Y., & Fong, B. (2022). A Study of Human-AI Symbiosis for Creative Work: Recent Developments and Future Directions in Deep Learning. *ACM Transactions on Multimidia Computing Communications and Applications*.

Malaspina, C. (2019). Pure information: On infinity and human nature in the technical object. *Culture, Theory and Critique*, *60*(3–4), 205–222.

Marcuse, H. (1964). *One-dimensional man: Studies in the ideology of advanced industrial society*. Beacon Press.

Marcuse, H. (2000). *An essay on liberation*. Beacon Pr.

Marx, K. (1993). *Grundrisse: Foundations of the critique of political economy (rough draft)* (M. Nicolaus, Trans.; 23. print).

McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press.

McCulloch, W. S. (2003a). Introductory Remarks to the Tenth Conference. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 687–688). Diaphanes. (Original work published 1953)

McCulloch, W. S. (2003b). Summary of the Points of Agreement Reached in the Previous Nine Conferences on Cybernetics. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 719–725). Diaphanes. (Original work published 1953)

McCulloch, W. S. (2004). The beginning of cybernetics. In C. Pias (Ed.), *Cybernetics/Kybernetik—The Macy Conferences 1946-1953: Essays and documents: Vol. II* (1. Aufl, pp. 345–360). Diaphanes. (Original work published 1974)

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

Mehta, I. (2003, April 18). *Elon Musk wants to develop TruthGPT, 'a maximum truth-seeking AI' | TechCrunch*. https://techcrunch.com/2023/04/18/elon-musk-wants-to-develop-truthgpt-a-maximum-truth-seeking-ai/

Mindell, D. A. (2002). *Between human and machine: Feedback, control, and computing before cybernetics*. The Johns Hopkins University Press.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, *49*(1), 8–30.

Minsky, M. (2006). The Emotion Machine: Commonsense Thinking. *Artificial Intelligence, and the Future of the Human Mind. Simon & Schuster*.

Minsky, M., & Papert, S. A. (1972). *Perceptrons: An introduction to computational geometry* (2. print. with corr). The MIT Press. (Original work published 1969)

Müller, V. C. (2021). Deep Opacity Undermines Data Protection and Explainable Artificial Intelligence. *Overcoming Opacity in Machine Learning*, 18.

Nadimpalli, M. (2017). Artificial intelligence risks and benefits. *International Journal of Innovative Research in Science, Engineering and Technology*, *6*(6).

Natale, S. (2021). *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press.

Ng, A. (2017, April 7). *Machine Learning by Professor Andrew Ng [Coursera]*. GitHub. https://github.com/blitz70/ML/blob/master/LectureAndrewNg.md

Noam Chomsky on AI: The Singularity is Science Fiction! (2013, October 5). Singularity Weblog. https://www.singularityweblog.com/noam-chomsky-the-singularity-is-science-fiction/

Orwell, G. (1949). *1984* (Nachdr.). Signet.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback* (arXiv:2203.02155). arXiv. http://arxiv.org/abs/2203.02155

Partnership on AI. (n.d.). Partnership on AI. Retrieved May 3, 2023, from https://partnershiponai.org/

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

Paz, B. J. (2020). Kai-Fu-Lee (2019): AI Superpowers—China, Silicon Valley and the New World Order. *AI & SOCIETY*, *35*(3), 771–772. https://doi.org/10.1007/s00146-020-00991-3

Pérez, J., Marinković, J., & Barceló, P. (2019). On the turing completeness of modern neural network architectures. *arXiv Preprint arXiv:1901.03429*. https://arxiv.org/abs/1901.03429

Perrigo, B. (2023, February 17). *Bing's AI Is Threatening Users. That's No Laughing Matter*. Time. https://time.com/6256529/bing-openai-chatgpt-danger-alignment/

Peter-Paul Verbeek. (2006). Materializing Morality: Design Ethics and Technological Mediation. *Science, Technology & Human Values*, *31*(3), 361–380.

Pias, C. (Ed.). (2003a). *Cybernetics - the Macy Conferences 1946-1953: The complete transactions* (First printing). Diaphanes.

Pias, C. (Ed.). (2003). Possible Mechanisms of Recall and Recognition. In *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 122–159). Diaphanes. (Original work published 1949)

Pias, C. (2003b). The Age of Cybernetics. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions, 2003* (First printing, pp. 11–26). Diaphanes.

Protocol sentence | Logic, Meaning & Truth | Britannica. (n.d.). Retrieved April 4, 2024, from https://www.britannica.com/topic/protocol-sentence

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Roberts, B. (n.d.). *Why You Should See Cats, Even If You Think It Looks Terrible | Cinemablend*. Retrieved June 3, 2024, from https://www.cinemablend.com/news/2487204/why-you-should-see-cats-even-if-you-think-it-looks-terrible

Rodolfo, K. (2000, January 3). *What is Homeostasis?* Scientific American. https://www.scientificamerican.com/article/what-is-homeostasis/

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, *10*(1), 18–24.

Rossi, F. (2016). *Artificial intelligence: Potential benefits and ethical considerations*.

Rouvroy, A., & Berns, T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation. *Reseaux*, *n° 177*(1), 163–196.

Rudolf Carnap > C. Inductive Logic (Stanford Encyclopedia of Philosophy). (n.d.). Retrieved April 4, 2024, from https://plato.stanford.edu/entries/carnap/inductive-logic.html

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, ed. DE Rumelhart and J. McClelland. Vol. 1. 1986. *Biometrika*, *71*, 599–607.

Russell, S. J., Norvig, P., Chang, M., Devlin, J., Dragan, A., Forsyth, D., Goodfellow, I., Malik, J., Mansinghka, V., Pearl, J., & Wooldridge, M. J. (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Samuel, A. L. (1974, December 11). *Pamela McCorduck Collection* (Transcripts Box 2 Folder 56) [Interview]. Carnegie Mellon University Archives.

Schell, H. (2023, April 30). *Which Jobs Will Be Most Impacted by ChatGPT?* Visual Capitalist. https://www.visualcapitalist.com/cp/which-jobs-artificial-intelligence-gpt-impact/

Science: The Thinking Machine. (1949, January 24). *Time*. https://content.time.com/time/magazine/article/0,9171,799721,00.html

Scott, D. (2014). *Gilbert Simondon's Psychic and collective individuation: A critical introduction and guide*. Edinburgh University Press.

Scott, R. (Director). (1982, June 25). *Blade Runner*. Warners Bros. Pictures.

Seibt, J., & Rodogno, R. (2019). Understanding emotions and their significance through social robots, and vice versa. *Techné: Research in Philosophy and Technology*, *23*(3), 257–269.

Sejnowski, T. J. (2018). *The deep learning revolution*. The MIT Press.

Shahriari, K., & Shahriari, M. (2017). IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, 197–201.

Shannon, C. (2003). Presentation of a Maze-Solving Machine. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 474–479). Diaphanes. (Original work published 1951)

Shannon, C. (2003). The Redundancy of English. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 248–272). Diaphanes. (Original work published 1950)

Shevlin, H., Vold, K., Crosby, M., & Halina, M. (2019). The limits of machine intelligence. *EMBO Reports*, *20*(10), e49177. https://doi.org/10.15252/embr.201949177

Shingler, B. (2023, May 6). Before worrying about AI's threat to humankind, here's what else Canada can do | CBC News. *CBC*. https://www.cbc.ca/news/canada/ai-canada-legislation-1.6833653

Shirodkar, N. (n.d.). *Maxi*. Retrieved June 3, 2024, from https://cathome.sg/cat?hash=2e8d4430db5011ed86d23da9d939d9f1

Shirodkar, N. (2023). *Infant killed by family's hybrid wolf-dog pet in US state*. https://pune.news/international/infant-killed-by-familys-hybrid-wolf-dog-pet-in-us-state-98272/

Siegelmann, H. T., & Sontag, E. D. (1992). On the computational power of neural nets. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 440–449. https://doi.org/10.1145/130385.130432

Simondon, G. (1989). *Du mode d'existence des objets techniques* (Ed. augm). Aubier. (Original work published 1958)

Simondon, G. (2005). *L'individuation à la lumière des notions de forme et d'information*. Millon. (Original work published 1964)

Simondon, G. (2016a). Cybernétique et philosophie. In *Sur la philosophie* (1 edition). Presses Universitaires de France. (Original work published 1953)

Simondon, G. (2016b). Épistémologie de la cybernétique. In *Sur la philosophie* (1 edition). Presses Universitaires de France. (Original work published 1953)

Simondon, G. (2016). *On the mode of existence of technical objects*. Univocal Pub. (Original work published 1958)

Simondon, G. (2020). *Individuation in Light of Notions of Form and Information* (T. Adkins, Trans.). Univ Of Minnesota Press. (Original work published 1964)

Sipser, M. (2013). *Introduction to the theory of computation* (Third edition, international edition). Cengage Learning.

Snow, C. P. (1993). *The two cultures* (Canto ed). Cambridge University Press. (Original work published 1959)

Socrates. (2013, October 5). *Noam Chomsky on AI: The Singularity is Science Fiction!* Singularity Weblog. https://www.singularityweblog.com/noam-chomsky-the-singularity-is-science-fiction/

Solomonoff, G. (2019). Ray Solomonoff and the Dartmouth summer research project in artificial intelligence. *No Date), Http://Raysolomonoff. Com/Dartmouth/Dartray. Pdf*.

Solomonoff, R. J. (1960). *A preliminary report on a general theory of inductive inference*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a510b16a624e512de69e4ede947d5989b44823b3

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control*, *7*(1), 1–22.

Solomonoff, R. J. (1975). Inductive Inference Theory-A Unified Approach to Problems in Pattern Recognition and Artificial Intelligence. *IJCAI*, 274–280.

Solomonoff, R. J. (1978). Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, *24*(4), 422–432.

Solomonoff, R. J. (1985). The time scale of artificial intelligence: Reflections on social effects. *Human Systems Management*, *5*(2), 149–153.

Solomonoff, R. J. (1997). The discovery of algorithmic probability. *Journal of Computer and System Sciences*, *55*(1), 73–88.

Solomonoff, R. J. (2006). Machine learning-past and future. *Dartmouth, NH, July*.

Solomonoff, R. J. (2011). Algorithmic probability–its discovery–its properties and application to strong AI. *Randomness Through Computation: Some Answers, More Questions*, 1–23.

Stiegler, B. (1998). *Technics and time, 1: The Fault of Epimetheus*. Stanford University Press.

Stiegler, B. (2010a). *For a new critique of political economy* (D. Ross, Trans.). Polity.

Stiegler, B. (2010b). *Taking care of youth and the generations*. Stanford Univ. Press.

Stiegler, B. (2011). *Technics and time, 3: Cinematic time and the question of malaise* (S. F. Barker, Trans.). Stanford University Press.

Stiegler, B. (2016). *Automatic society: The future of work*. Polity Press.

Stiegler, B. (2020). *Nanjing Lectures (2016-2019)* (D. Ross, Ed.). Open Humanities Press.

Stray, J. (2020). Aligning AI Optimization to Community Well-Being. *International Journal of Community Well-Being*, *3*(4), 443–463. https://doi.org/10.1007/s42413-020-00086-3

Stroud, J. (2003). The Psychological Moment in Perception. In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 41–65). Diaphanes. (Original work published 1949)

Sundin, M. (2023, May 9). *❌ No, 50% of AI researchers don't believe there is a 10% risk that AI will kill us*. Warp News. https://www.warpnews.org/artificial-intelligence/no-50-of-ai-researchers-dont-believe-there-is-a-10-risk-that-ai-will-kill-us/

Supersaturation. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Supersaturation&oldid=1084840065

Symbiosis Definition & Meaning—Merriam-Webster. (n.d.). Retrieved May 25, 2023, from https://www.merriam-webster.com/dictionary/symbiosis

Symbolic artificial intelligence. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Symbolic_artificial_intelligence&oldid=1112958138

Tegmark, M. (2023, April 25). *The "Don't Look Up" Thinking That Could Doom Us With AI | Time*. https://time.com/6273743/thinking-that-could-doom-us-with-ai/?ref=warpnews.org

The Declaration—Montreal Responsible AI. (2017). https://www.montrealdeclaration-responsibleai.com/the-declaration

Turing, A. (1948). *Intelligent machinery*.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *J. of Math*, *58*(345–363), 5.

Turner, F. (2008). *From counterculture to cyberculture: Stewart Brand, the Whole Earth Network, and the rise of digital utopianism* (1. paperback ed). Univ. of Chicago Pr.

Urone, P. P., Hinrichs, R., OpenStax College, Open Textbook Library, & OpenStax (Nonprofit organization). (2020). *Physics: High school*. https://openstax.org/details/books/physics

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. *Science Fiction Criticism: An Anthology of Essential Writings*, 352–363.

von Foerster, H., Mead, M., & Teuber, H. L. (2003). A Note by the Editors (Transactions of the Eighth Conference). In C. Pias (Ed.), *Cybernetics—The Macy Conferences 1946-1953: The complete transactions* (First printing, pp. 341–348). Diaphanes. (Original work published 1951)

Watson, D. S. (2021). No Explanation without Inference. *Overcoming Opacity in Machine Learning*, 9.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models* (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682

Weinstein, L. (Director). (1997). *The War Symphonies: Shostakovich against Stalin*. Decca.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. Freeman.

Wiener, N. (1989). *I am a mathematician: The later life of prodigy ; an autobiographical account of the mature years and career of Norbert Wiener and a contination of the account of his childhood in ex-prodigy* (6. pr). M.I.T. Pr. (Original work published 1956)

Wiener, N. (2007). *Cybernetics or control and communication in the animal and the machine* (2. ed., 14. print). MIT Press. (Original work published 1948)

Wiener, N. (2018). *Norbert Wiener—A Life in Cybernetics: Ex-Prodigy: My Childhood and Youth and I Am a Mathematician: The Later Life of a Prodigy*. https://doi.org/10.7551/mitpress/11597.001.0001 (Original work published 1956)

Wiggins, C., & Jones, M. L. (2023). *How data happened: A history from the age of reason to the age of algorithms* (First edition). W. W. Norton & Company.

Winner, L. (1978). *Autonomous technology: Technics-out-of-control as a theme in political thought*. Mit Press.

Winner, L. (1993). Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology. *Science, Technology, & Human Values*, *18*(3), 362–378.

Winner, L. (2010). *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press.

Winner, L., Feenberg, A., & Nielsen, T. H. (1997). *Technology and Democracy: Technology in the Public Sphere*. Oslo: Center for Technology and Culture.

Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Ablex Pub. Corp.

Winston, P. H. (2000). *Artificial intelligence* (3. ed., repr. with corr., 13. [pr.]). Addison-Wesley.

Yeasmin, S. (2019). Benefits of artificial intelligence in medicine. *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, 1–6.

Yudkowsky, E. (2023, March 29). The Open Letter on AI Doesn't Go Far Enough: We Need to Shut it All Down. *Time*. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/

Zahedi, Z., & Kambhampati, S. (2021). Human-AI symbiosis: A survey of current approaches. *arXiv Preprint arXiv:2103.09990*.

Zahedi, Z., Sreedharan, S., & Kambhampati, S. (2022). A Mental-Model Centric Landscape of Human-AI Symbiosis. *arXiv Preprint arXiv:2202.09447*.

Zednik, C., & Boelsen, H. (2021). Preface: Overcoming Opacity in Machine Learning. *Overcoming Opacity in Machine Learning*, 1–2.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, *32*(4), 661–683.

Zhang, X., Wei, X., Ou, C. X., Caron, E., Zhu, H., & Xiong, H. (2022). From human-AI confrontation to human-AI symbiosis in society 5.0: Transformation challenges and mechanisms. *IT Professional*, *24*(3), 43–51.

Zielinski, S. (2020). *Let's learn about domestic cats*. https://www.snexplores.org/article/lets-learn-about-domestic-cats

Zimmerman, M. E. (2016). Heidegger on Techno-Posthumanism: Revolt against Finitude, or Doing What Comes “Naturally”? In J. B. Hurlbut & H. Tirosh-Samuelson (Eds.), *Perfecting Human Futures: Transhuman visions and technological imaginations* (pp. 121–140). Springer Fachmedien Wiesbaden GmbH.

Appendix: Language Model in Machine Learning

This Appendix presents an overview of how human languages can be represented as computer models. Large language models like GPT (Generative Pre-trained Transformer) uses the transformer model originally designed for natural language processing tasks such as language translation. It consists of an encoder and a decoder. In automatic language translation, the encoder would first encode a sentence in the source language into a meaningful representation. The decoder would then decode this meaningful representation into a sentence in the target language.

Here is an abridged explanation[177] of how the encoder of the transformer model works and what it uses a neural network for. To encode the sentence "We are going to send her a card" into some meaningful representation, the sentence is first divided into tokens: <SOS> (a special token for the start of a sentence), "We," "are," "going," "to," "send," "her," "a," and "card," and <EOS> (a special token for the end of a sentence). Each token is converted into a token embedding (see Figure A1). To see what token embedding means, let us focus on the meaning of the token "card." This card can mean a postal card, a bank card, or a game card. Suppose that, without considering the context of how the word appears in a sentence, there is an 85% chance that the card is a bank card, 60% chance that it is a postal card, and 15% chance that it is a game card. The token embedding for the token "card" would be the vector [0.85, 0.6, 0.15]. In other words, the token embedding contains information about the possible contexts in which the word appears. In GPT-3, the dimensionality of a token embedding is 768. Each token embedding has 768 possible contexts and is a vector with 768 entries. The token embedding for "card" may then look something like [0.002, 0.003, 0.001, 0.85, … 0.6, … 0.15, …], in which case the default probability of a "bank" context would be the number in the fourth entry of the vector. Each token embedding is then further encoded with the position of the token in the original sentence.[178] In other words, each token is converted

[177] I will take out details such as the "attention layers" to simplify my explanation.

[178] The positioning encoding uses a set of sines and cosines at different frequencies across the sequence. For example, with "card" in the 9th position, the token embedding encoded with the 9th position would become [0.002+$f_0(9)$, 0.003+$f_1(9)$, 0.001+$f_2(9)$, 0.85+$f_3(9)$, …] where $f_i(pos)$ is some sine or cosine wave function. For our purpose, we only need to know that the numbers in the token embedding is modified by positioning encoding and that the position can be decoded at a later stage.

to a positional token embedding, which is a vector that contains information on the default context and its position in a sentence. This encoded token embedding is then passed sequentially to a trained neural network (see Figure A2).

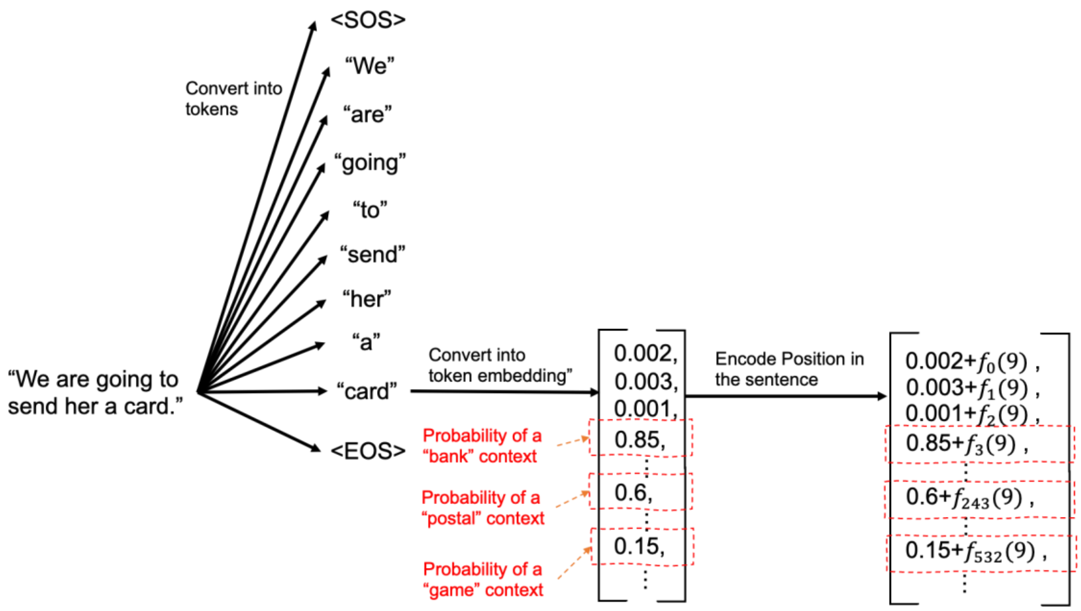


Figure A1: Converting a sentence into inputs to the neural network.

By the time this neural network operates on the positional token embedding for "card," it would have already done the same for the token "send." The awareness of "send" as a token in the same sentence as "card" makes the neural network realize that "card" is very likely to be a postal card. The neural network would therefore tune the probabilities about the probable contexts for the token "card." If the positional encoding is not considered for simplicity's sake, the adjusted token embedding would look something like [… 0.2, … 0.95, … 0.1, …], which indicates that the card has a 20% chance of being a bank card, 95% chance of being a post card, and 10% chance of being a game card (see Figure A3). Once the <EOS> token is processed, the neural network would then sequentially output the adjusted token embeddings (along with the encoded position) that correspond to the input token embeddings. Note that the neural network does not need explicit knowledge about the individual tokens. Instead, it operates on the token embeddings, which capture the semantic or contextual information associated with each token. In other words, the token embeddings serve as numerical representations that encode the meaning or characteristics of the tokens. The neural network would process these numerical representations to capture the relationships and contextual information among the tokens. It then expresses such semantic information as a sequence of output token embeddings.

There are many ways to train an encoder's neural network. One popular method is to use the Masked Language Model. For instance, for the masked sentence "The [MASK1] brown fox [MASK2] over the lazy dog," the neural network is trained to generate the outputs "quick" for [MASK1] and "jumped" for [MASK2]. This design for training a neural network comes with the following assumption: If a neural network can do well in such fill-in-the-blank problems, it must have extrapolated relationships and contextual information from the masked sentence. In this training set-up, the encoded numerical representation output can be converted into probabilities of tokens (see Figure A4), and the neural network can be trained to minimize the deviation between these probabilities and the correct meaning for the masked token.

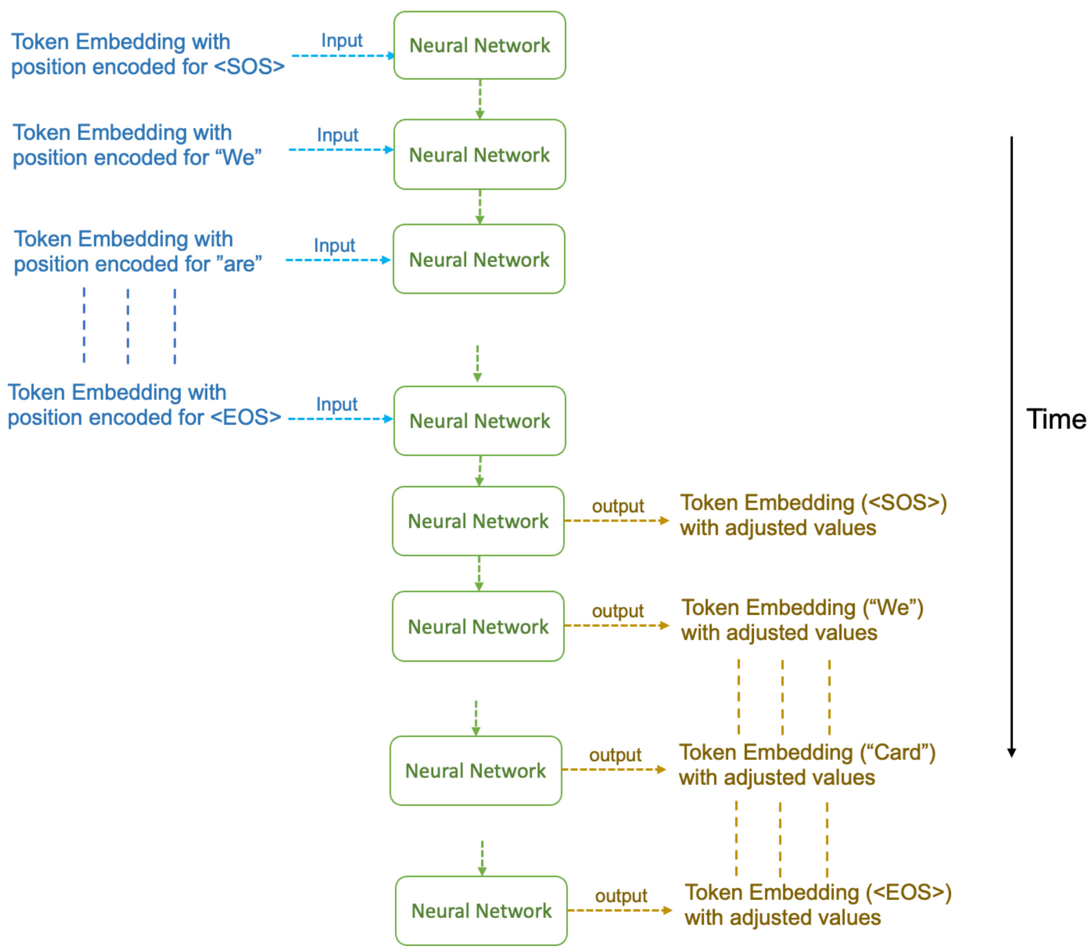


Figure A2: Token Embedding with position encoded are input sequentially to the neural network in the encoder of the transformer.

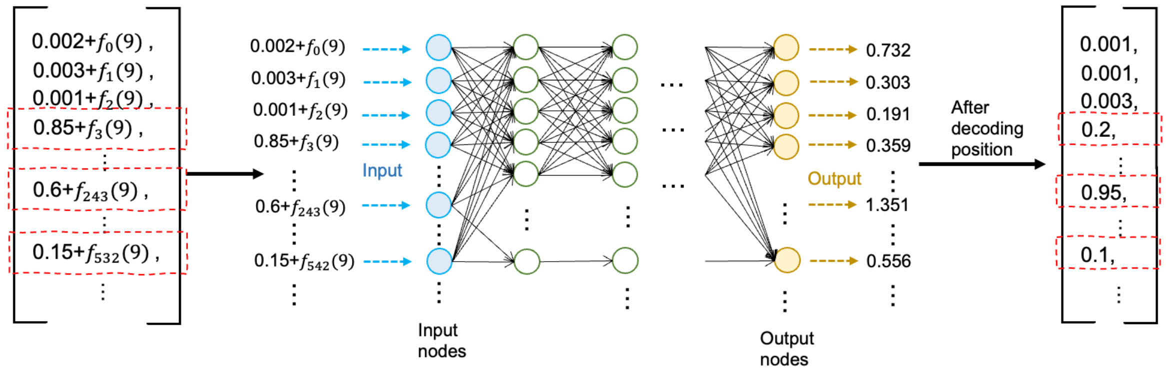


Figure A3: Neural Network in Encoder

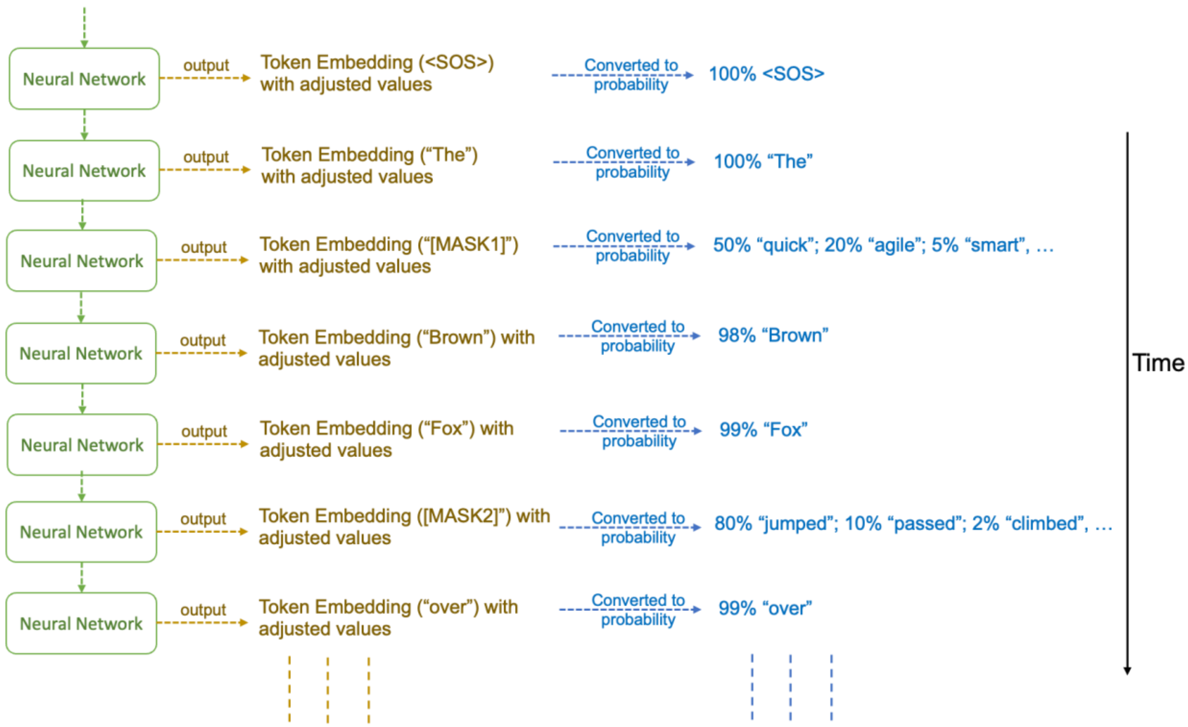


Figure A4: Masked Language Model