2022

Contextualizing Artificial Intelligence: The History, Values, and Epistemology of Technology in the Philosophy of Science

Christopher Grimsley
University of Kentucky, christopher.grimsley@outlook.com
Digital Object Identifier: https://doi.org/10.13023/etd.2022.199

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Christopher Grimsley, Student

Dr. Julia Bursten, Major Professor

Dr. Tim Sundell, Director of Graduate Studies

Contextualizing Artificial Intelligence: The History, Values, and Epistemology of Technology in the Philosophy of Science

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Christopher M. Grimsley
Lexington, Kentucky

Director: Dr. Julia Bursten, Associate Professor of Philosophy
Lexington, Kentucky
2022

ABSTRACT OF DISSERTATION

Contextualizing Artificial Intelligence: The History, Values, and Epistemology of Technology in the Philosophy of Science

Artificial intelligence (AI) and other advanced technologies pose new questions for philosophers of science regarding epistemology, science and values, and the history of science. I will address these issues across three essays in this dissertation. The first essay concerns epistemic problems that emerge with existing accounts of scientific explanation when they are applied to deep neural networks (DNNs). Causal explanations in particular, which appear at first to be well suited to the task of explaining DNNs, fail to provide any such explanation. The second essay will explore bias in systems of automated decision-making, and the role of various conceptions of objectivity in either reinforcing or mitigating bias. I focus on conceptions of objectivity common in social epistemology and the feminist philosophy of science. The third essay probes the history of the development of 20th century telecommunications technology and the relationship between formal and informal systems of scientific knowledge production. Inquiring into the role that early phone and computer hackers played in the scientific developments of those technologies, I untangle the messy web of relationships between various groups that had a lasting impact on this history while engaging in a conceptual analysis of "hacking" and "hackers."

KEYWORDS: philosophy of science, science and values, artificial intelligence, explanation, bias, hacking

Christopher M. Grimsley

1st May, 2022

Contextualizing Artificial Intelligence: The History, Values, and Epistemology of Technology in the Philosophy of Science

By
Christopher M. Grimsley

Dr. Julia Bursten
Director of Dissertation

Dr. Tim Sundell
Director of Graduate Studies

1st May, 2022
Date

ACKNOWLEDGMENTS

I am so very thankful to so many people for helping to make this dissertation possible. First and foremost I would like to thank my advisor, Dr Julia Bursten, for her excellent help and guidance through the dissertation process. I am also very thankful to the excellent group of scholars at the University of Kentucky who are currently serving or have served on my committee: Dr Judy Goldsmith, Dr Meg Wallace, Dr Natalie Nenadic, Dr Clare Batty, Dr Tim Sundell, and Dr Michael Baker.

Thank you to those outside of my committee who have directly or indirectly provided feedback on parts of this dissertation or otherwise helped me to shape my ideas about it: Dr Angela Potochnik, Dr Collin Rice, Dr Colin Allen, Dr Robert Scharff, Dr Kelle Dhein, and Dr Eric Thomas Weber.

I would also like to thank the many educators who helped to shape the course of my academic and professional life: Richard Haskins, Judy Carter, Windy Spiridigliozzi, Eric Eiswert, Rich Hambor, Dr Gary Blankenburg, Dr Gerald Snelson, Dr Joy Kroeger-Mappes, Dr Jean-Marie Makang, Rachel Hoover, Dr Todd Rosa, Cynthia Crable, Dr Amy Branam-Armiento, and Dr Kevin Knott.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1 Causal and Non-Causal Explanations of Artificial Intelligence

1.1 Problems with Existing Accounts of Explanation

This chapter will engage in a deep analysis of a serious open problem in explainable artificial intelligence (XAI) right now, the simultaneous existence of the widespread acknowledgement of the need for XAI on the one hand, and the lack of a cohesive account of scientific explanation which fits AI on the other. This problem has several consequences, chief among them is that the meaning of the phrase "explainable AI" changes depending on who is speaking or writing, and without an agreed-upon definition of this phrase, it will remain impossible to construct explainable AI. While many researchers are working toward what they believe to be the same end goal, they haven't all defined the problem in the same way, and I don't believe we can make significant progress toward XAI until we have an account of scientific explanation that fits AI.

AI is not a technology of the future – it is in constant use every day, and it impacts nearly everyone who uses the Internet or a smart phone; it impacts activities as mundane as grocery shopping[1] or as important as national elections.[2] AI has been used to make decisions about granting or denying loan applications,[3] bail,[4] and parole.[5] It is at the core of facial recognition systems that have been used to identify (or sometimes misidentify) criminal suspects in the US,[6] and Uyghur Muslims in China.[7] There are enormous social consequences to the use of the AI systems that

[1]Murshed et al., *Hazard Detection in Supermarkets using Deep Learning on the Edge*.

[2]Kaiser, *Targeted: The Cambridge Analytica Whistleblower's Inside Story of How Big Data, Trump, and Facebook Broke Democracy and How It Can Happen Again*, p. 161.

[3]Fuster et al., "Predictably unequal? the effects of machine learning on credit markets".

[4]Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead".

[5]Khademi and Honavar, "Algorithmic Bias in Recidivism Prediction: A Causal Perspective".

[6]Hill, *Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match*.

[7]Mozur, *One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority*.

already exist. The people who are impacted by these practices deserve an explanation.

Practices like redlining in real estate or racial profiling in policing are both illegal and unethical, so it is important to be certain that they are not simply being carried forward under the guise of some kind of supposedly neutral and value free algorithmic decision making. Maciej Cegłowski described machine learning as "money laundering for bias,"[8] meaning that AI often gives the false sense that decisions are being made free from human-introduced bias. Unfortunately what really occurs is the bias that already pervades society gets baked into AI, but then in the process becomes less obvious. So institutions end up still making the same bad and often racist choices, but it's significantly harder to track down the source of the problem. One solution is to have a robust account of explanation which is substantial enough to adequately explain AI — something which does not currently exist. The goal of this chapter is to provide a starting point for the development of such an account of explanation.

'Machine learning,'[9] an increasingly common form of AI, is a broad term that describes programs that accept and process unexpected input data without being explicitly programmed to do so. One of the more common contemporary approaches to machine learning is the neural network. Neural networks function analogously to the behavior of biological brains by linking input and output together via various intermediary nodes in a network. Each node is called a 'neuron', hence 'neural network'. Neural networks contain multiple layers of neurons including an input layer, an output layer, and one or more 'hidden layers' between the input and output. Each neuron is a node in the neural network, and the neurons in each layer have edges which connect to nodes in adjacent layers. Each node has a function which processes its input and produces its output. The neurons also have a weight, with higher-weighted nodes having more control over the production of the final output

[8] Ceglowski, *Privacy Rights and Data Collection in a Digital Economy*.

[9] for a more comprehensive overview, see Buckner ("Deep Learning: A Philosophical Introduction").

than lower-weighted nodes.

Deep neural networks (DNNs) are neural networks with a high degree of complexity, typically containing more than three hidden layers. DNNs produce a complex, often non-interpretable model that is used in decision or classification tasks. In what is called 'supervised learning,' a 'trained model' is created by providing labeled datasets to the DNN, which iterates over the labeled data and builds a model capable of making the correct decision or classification given novel data. In other words, the deep neural model is built with the deep neural network. DNNs and the models they produce are both in need of explanation.

Though building a suitable account of explanation for AI, especially in the form of DNNs, is a difficult task, there is already a deep literature on explanation in the philosophy of science, which can serve as a solid starting point for the creation of a list of desired qualities for XAI. Computer scientists have broken interesting technological ground in the development of techniques which serve to explain particular AI systems, but without more engagement with the philosophy of science this problem can't be fully solved. The problem of XAI is more than a just mathematical, computational, or technological problem; it is a philosophical problem.

A survey of various accounts of scientific explanation in the philosophy of science going back to the 1940s reveals the scope of the problem. AI, particularly in the form of DNNs, is difficult to explain by any account, even as it produces sensible and coherent output. DNNs are among the most complex neural networks; they are described as "deep" due to the increased number of layers in the network compared to other neural networks. AI, especially in the form of the DNN, is not adequately explained by any of the most frequently cited accounts of explanation, including those offered by Hempel,[10] Hempel and Oppenheim,[11] Salmon (1971),[12] Salmon (1984),[13]

[10]Hempel, "The Function of General Laws in History".
[11]Hempel and Oppenheim, "Studies in the Logic of Explanation".
[12]Salmon, *Statistical Explanation and Statistical Relevance*.
[13]Salmon, *Scientific Explanation and the Causal Structure of the World*.

Kitcher,[14] and Woodward.[15] One way of thinking about the problem is to say that AI currently lacks a satisfactory scientific explanation under any account in the philosophy of science, but one corollary of this is that AI serves as a counterexample to every existing account of scientific explanation in the philosophy of science.

Hempel and Oppenheim[16] argued that an explanation for a given phenomenon is derivable from the combination of antecedent conditions and general laws, and an explanation takes the form of a deductive argument. The problem with this structure of explanation when applied to AI is that we can think of the input to a DNN as the explanans, the structure of the neural network as general laws, and the output of the network as the explanandum. Though the deductive-nomological (DN) model appears to fit DNNs very well, it still does not explain them because despite having all of the relevant components of a DN explanation, a convincing explanation for many AI systems remains elusive. This is to say that XAI does not simply emerge in the presence of both the full explanandum and explanans under the DN model. The DN model is therefore inadequate as a basis for XAI.

The statistical relevance (SR) model[17] would appear at first to be a better candidate for explaining DNNs since neural networks rely on principles of statistics. A major problem in applying the SR model to DNNs centers on causality. Many of the problems to which AI is applied are causal questions. A statistical relationship between two variables does not imply a causal relationship. This is a problem with the way AI is used currently and it results in severely biased AI tools. In 2018, for instance, Amazon had to suddenly decommission an AI tool for screening job applicants because, using statistical methods, it determined that the only qualified candidates were men.[18] Decades of data on hiring decisions showed that men are more likely

[14]Kitcher, "Explanatory Unification and the Causal Structure of the World".
[15]Woodward, *Making Things Happen: A Theory of Causal Explanation*.
[16]Hempel and Oppenheim, "Studies in the Logic of Explanation".
[17]Salmon, *Statistical Explanation and Statistical Relevance*.
[18]Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*.

to be promoted to top positions in large corporations, but of course those decisions demonstrate a problem with bias in our society, not a truth about what types of people make the best job candidates.

Causal models, such as those of Wesley Salmon[19] and later adaptations from James Woodward[20] also fail to explain DNNs. The problem, plainly put, is that the complexity of DNNs prevents the creation of a complete account of causal relations between the relevant parts of the system. Without knowledge of which parts of the system are causally related, and more importantly which causal relations are relevant to the output, the use of causal explanations for AI will fail.

The potential for the use of models as explanations has been discussed by Bokulich,[21] Batterman and Rice,[22] Rohwer and Rice,[23] and Morrison[24] among others. Model explanations are an exciting possibility for DNNs because DNNs produce models which are used in decision and classification tasks. If models can serve as explanations, the explanation for DNNs could be found in the trained models (referred to as deep neural models). The model may be explanatory even if its structure does not correlate closely with the world — that is to say that it is an explanatory fiction or essential idealization — as long as there is an appropriate relationship between the relevant features of the two. Many, including Bokulich,[25] Rice,[26] and Potochnik[27] have argued that fiction can play a key role in understanding. One major problem with this approach is that, with the types of explanatory models discussed in the philosophy of science literature, the explanatory model either itself requires no explanation, or can be explained much more simply than can the phenomenon being

[19]Salmon, *Scientific Explanation and the Causal Structure of the World*.
[20]Woodward, *Making Things Happen: A Theory of Causal Explanation*.
[21]Bokulich, "How scientific models can explain".
[22]Batterman and C. C. Rice, "Minimal Model Explanations".
[23]Rohwer and C. Rice, "How are Models and Explanations Related?"
[24]Morrison, *Reconstructing Reality: Models, Mathematics, and Simulations*.
[25]Bokulich, "Distinguishing Explanatory from Nonexplanatory Fictions".
[26]C. Rice, "Moving Beyond Causes: Optimality Models and Scientific Explanation"; C. Rice, "Idealized Models, Holistic Distortions, and Universality".
[27]A. Potochnik, *Idealization and the Aims of Science*.

modeled. When a model of a target system is said to explain the target system, this is usually at least partially a function of the model being easier to understand than the target system. In these circumstances a recursive explanation requirement is avoided because the target system is explained by the model, and the model requires no explanation. This is not so with DNNs — an explanation is required both of the target system and the model, meaning that the model alone cannot be explanatory. This may highlight issues similar to those identified by Potochnik[28] related to optimality models and epistemic interdependence. There is a problem if the explanation of the deep neural network can be found in the deep neural model, because the trained model is as difficult to understand or explain as the DNN itself. This pushes the need for an explanation up one level, but does not eliminate it. The explanation of the trained model will suffer from the same problems as the explanation of the DNN. An explanation of the DNN that does not also explain the model (which is ultimately responsible for decision and classification tasks) is not enough. It is not just the DNN which requires an explanation, but the DNN and the model it produces.

1.2 The Need for Explainable Artificial Intelligence

Artificial intelligence (AI) is increasingly being used to make high-stakes decisions, often under questionable circumstances that indicate the presence of racial or gender bias, including granting or denying loan applications,[29] deciding which prisoners are eligible for parole,[30] and diagnosing mental health disorders.[31] If AI is used to make these decisions — especially if these decisions appear to have reinforced biases present elsewhere in society — understanding how the algorithm made the decision is

[28] Angela Potochnik, "Optimality Modeling and Explanatory Generality"; Angela Potochnik, "Explanatory Independence and Epistemic Interdependence: A Case Study of the Optimality Approach".

[29] Fuster et al., "Predictably unequal? the effects of machine learning on credit markets".

[30] Khademi and Honavar, "Algorithmic Bias in Recidivism Prediction: A Causal Perspective".

[31] Bennett and Keyes, "What is the Point of Fairness? Disability, AI and The Complexity of Justice".

essential. Absent explanation, arbitrary or biased decisions may go unchecked, so it is incumbent upon computer scientists to generate good explanations of the decision process happening inside AIs.

Computer scientists have recognized this problem and are actively engaged in ongoing efforts to develop new approaches to explainable AI. However, many of their strategies haphazardly employ a mix of causal, psychological, and counterfactual approaches to explanation. This fails to generate a theoretically-grounded conception of explanation, and, as I show below, it consequently generates roadblocks in providing adequate explanations in a number of cases. This is what I refer to as the explainability problem — the widespread recognition on the part of computer scientists and others that the development of explainable AI is paramount, along with the simultaneous imprecision with which the term "explainable AI" is used such that there is little to no agreement on which set of criteria serve as necessary or sufficient conditions for explainability. A potential solution can be found in the ways in which explanation is conceptualized within the context of AI. The philosophy of science is uniquely positioned to take on this problem and offer solutions by examining the meaning of scientific explanation and developing an account of explanation which adequately explains AI.

In this chapter, I use the history of the discourse in the philosophy of science around scientific explanation to give a theoretical underpinning to the notion of explanation in XAI. I highlight the ways in which that discourse can inform a broader understanding of explainable AI such that computer scientists are better able to build explainable systems. Scientific explanations, being situated within a broader discourse and literature in the history and philosophy of science, are more rigorously defined and more precisely articulated, which confers a distinct advantage over other forms of explanation when it comes to the development of XAI.

In the computer science literature, conversations around XAI are often paired with

conversations about "interpretable" AI. The lack of semantic consistency in the use of the term "explainable" mirrors an analogous problem surrounding interpretable AI, a similar, though distinct goal in AI development. Since this chapter primarily concerns explainability rather than interpretability I will take the term "interpretable" to mean that it is possible, given a particular AI or algorithmic system, to produce a full account of the transformation of the input to the output with a complete description of every function at each step along the way, including all relevant causal relations.

It may seem that interpretability is a sufficient condition for explainability, but I disagree. Interpretable algorithms are not necessarily explainable in cases where the transformation of the input to the output can be tracked programmatically but is too complex for any human to understand.

In recent years, advancements in the state of the art in AI have been achieved by increasing complexity. This process of increasing systemic complexity makes explanation harder. Within an AI system, the explanandum is often the output of the decision or classification task that the AI was designed for. Generally, a more complex explanandum will require more complex explanantia, thus as the process of generating this output becomes more complex, generating the explanantia becomes more difficult.

There have been attempts to create programmatically generated explanations of decision and classification tasks of particular AIs,[32,33,34,35] but in order to serve as XAI, these automated explanation programs must be generalizable to new generations of AI. As AI researchers continue to expand upon the state of the art in AI, they must also expand upon the state of the art in explanation. Because explanation through merely technological means is always lagging behind the complexity of the

[32]Kumar and Talukdar, *NILE : Natural Language Inference with Faithful Natural Language Explanations*.

[33]Puri et al., *Explain Your Move: Understanding Agent Actions Using Specific and Relevant Feature Attribution*.

[34]Darwiche and Hirth, *On The Reasons Behind Decisions*.

[35]Narang et al., *WT5?! Training Text-to-Text Models to Explain their Predictions*.

networks that are in need of an explanation, it is reasonable to conclude that the solution to this problem cannot be solved with more AI: the problem is conceptual, not technological.

In this chapter I argue that recent attempts by computer scientists to develop XAI fail because they do not employ a theoretically-grounded concept of explanation. Further, I show that it is necessary to employ non-causal accounts of explanation in order to solve the problem of explainability in AI. I begin with a brief overview of the aspects of AI that are relevant to my argument. Then I discuss two existing methods for developing XAI: one causal, and one non-causal. I demonstrate why each approach fails to generate a satisfactory explanation, then I propose alternative non-causal possibilities and explore the viability of each. I conclude that existing approaches to both causal and non-causal explanation fail to fit the needs of XAI, though of the two approaches, non-causal accounts hold greater promise.

1.3 The Current Landscape: Two Case Studies

Computer scientists have made use of two contrasting strategies in order to develop XAI. Most researchers attempting to build explainable DNNs appear to prefer causal forms of explanation,[36] however some have attempted to develop non-causally explainable DNNs. I present instances of each approach and discuss their relationships to the explanation literature in the philosophy of science.

1.3.1 Case Study One: "Rationalizations"

One approach to XAI is to develop algorithms that produce patterns of explanantia that imitate human reasoning. This is analogous to chatbots that imitate human

[36]See for example Yang et al. ("Who Did What: Editor Role Identification in Wikipedia."), Jain and Wallace ("Attention is not Explanation"), Khademi and Honavar ("Algorithmic Bias in Recidivism Prediction: A Causal Perspective"), and Sharma, Henderson, and J. Ghosh ("CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models")

texting patterns. For instance, Harrison et al.[37] use two AIs. The first is a typical game-playing AI which simply plays the classic video game Frogger. The second AI[38] is designed to explain the in-game actions of the game-playing AI. The explanations are generated by translating internal game state data to natural-language approximations of human-supplied explanations. In order to accomplish this, the research team recorded human subjects playing Frogger, then periodically paused the game and asked the subjects to verbally explain an action that they recently took. The human responses, combined with game state data were used to train an Encoder–Decoder network. This network, having access both to the human-supplied explanations of in-game actions and the game state corresponding to the in-game event that was explained is able to generate plausible human-like explanations of game events encountered by the game-playing AI.

Importantly, the explanation-generating Encoder–Decoder network was not generating veridical statements about the internal state of the game-playing AI, but was generating a unique natural-language statement based on data gathered from human players when in similar in-game situations. This approach generates psychologically satisfying explanations of AI behavior. Essentially, the explanation-generating Encoder–Decoder network was given many examples of game events (as game state data) along with many examples of human-supplied explanations of those game events, and used this to generate explanations of in-game actions taken by the AI comprised of "synthetic sentences grounded in natural language."[39] Because the generated explanations are only meant to approximate human-supplied explanations of similar situations, a trade-off is made between accurately reporting internal AI

[37]Harrison, Ehsan, and Riedl, "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations".

[38]The architecture used was an Encoder–Decoder network, a type of Recurrent Neural Network (RNN)(Harrison, Ehsan, and Riedl, "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations", p. 3)

[39]Harrison, Ehsan, and Riedl, "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations", p. 4.

states and psychologically satisfying explanations. The authors accept this trade-off in order to obtain quickly-generated and human-like explanations. The authors write that "rationalization is fast, sacrificing absolute accuracy for real-time response".[40]

The explanation-generating Encoder–Decoder network does not supply a veridical explanation of the decision making process used by the game-playing AI. Instead it produces statements that approximate human-generated explanations when faced with similar in-game circumstances. It was not the intent of the researchers to develop accurate explanations of what the AI is actually doing, but instead they wanted to reconstruct how a person in a similar situation might have explained their actions. While it may be intuitive to a human that, for example, a playable character moved left to avoid a game-ending collision, this may not actually have anything to do with the process happening within the game-playing AI's logic. Another much deeper problem with this model is that, since the explanation of one AI is itself generated by a different, independent AI, there is now a need for an explanation of the explanation. If one black-box system is explained by appealing to a second black-box system, nothing has actually been explained. The number of phenomena in need of explanation has actually increased.

It should be noted that the goal of the rationalization approach to explainable AI is not to provide deep, correct, technical explanations, but to provide explanations that satisfy the human members of teams that use AI in their work. If humans depend on the use of AI for a critical task, it is important that a sense of trust in that AI is maintained. One goal of the research of Harrison et al.[41] is to provide explanations that reassure human operators of AI that the AI had a good reason for doing an action that may appear to a human to be questionable. In some cases this may mean that the AI only needs to be able to communicate that a good reason for

[40]Harrison, Ehsan, and Riedl, "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations", p. 1.

[41]Harrison, Ehsan, and Riedl, "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations".

a particular action exists, i.e. to articulate a *how-possibly* explanation, rather than communicating the right reason for the action, i.e. a *how-actually* explanation.

Rationalizations are an attempt to deal with the problems associated with the lack of XAI without actually solving them. The authors endorse the view that, when it comes to AI, we must choose between fast, intuitive, human-understandable explanations, and technically correct explanations. Rationalizations do not attempt to provide explanations, but instead provide fictional statements that sound like plausible explanations.

1.3.2 Why Rationalizations are not Explanations

Rationalizations represent only one attempt to build non-causal XAI, but this attempt leaves much to be desired from the standpoint of scientific explanation. Rationalizations are explicitly non-veridical. While this is a problem for rationalizations, it isn't necessarily a problem for all explanations. In fact, fiction often serves a role in scientific explanation. Many, including Bokulich,[42] Potochnik,[43] and Rice[44] have argued that fiction can play a key role in understanding. Rationalizations differ from fictions in other models. It is not always necessary that an explanation avoid all use of fiction, but if a fiction is to be part of a successful explanation it must represent the correct relevance relations of the target system. That is to say that fictions can be explanatory in cases where they highlight the actual relationships among the elements of real world systems. If the understanding that an explanation helps to foster is not in any sense an understanding of a true state of affairs, then the purported explanation has not contributed to epistemic success, and is not actually explanatory.

Bokulich[45] distinguishes between explanatory and non-explanatory fictions, and

[42]Bokulich, "How scientific models can explain"; Bokulich, "Distinguishing Explanatory from Nonexplanatory Fictions".

[43]A. Potochnik, *Idealization and the Aims of Science*.

[44]C. Rice, "Idealized Models, Holistic Distortions, and Universality".

[45]Bokulich, "Distinguishing Explanatory from Nonexplanatory Fictions".

acknowledges that it is reasonable to include fictional elements in a scientific model which nevertheless remains explanatory. The existence of a fiction within an explanation does not itself prohibit that explanation from being successful, but only under certain conditions. An example of a non-explanatory fiction provided by Bokulich[46] is Ptolemaic astronomy. In order for the geocentric model to work, it must include planetary epicycles to explain the apparent retrograde motion of the planets as observed from Earth. Even though the geocentric model can correctly predict the position of the planets, the fiction in this case is non-explanatory, not simply because it is false, but because of relevance relations within the explanation. According to Bokulich, "only those fictions that are an adequate representation of the relevant features of the world are admitted into the scientist's explanatory store".[47] Planetary epicycles are not relevant in the explanation of retrograde motion, not because they are fictional, but because they don't capture the actual relationships between real objects in the world, even if only fictionally.

A related set of problems in philosophy center on the metaphysical relationships between utterances meant to be taken at face value and the states of the world those utterances are meant to represent. Similar to the rationalization approach to XAI described above, fictionalism[48] regards many truth claims as aiming, not at literal truth, but as a truth-approximating fiction. Fictionalism has been convincingly argued for with regard to numbers,[49] scientific discourse,[50] and everyday objects.[51] Kroon, McKeown-Green, and Brock explain that a problem emerges when one makes claims such as "the sun rose at 6:05 this morning"[52] because, while this claim is not literally true[53], it is uttered with the expectation that it will be taken at face value.

[46]Bokulich, "Distinguishing Explanatory from Nonexplanatory Fictions".
[47]Bokulich, "Distinguishing Explanatory from Nonexplanatory Fictions", p. 734.
[48]Eklund, "Fictionalism".
[49]Field, *Science without Numbers*.
[50]Fraassen, Press, and Van Fraassen, *The Scientific Image*.
[51]Inwagen, *Material Beings*.
[52]Kroon, McKeown-Green, and Brock, *A Critical Introduction to Fictionalism*, p. 1.
[53]The sun does not literally rise or set, but only appears to do so from particular vantage points

The assertion, while indirectly pointing to something that is true, also embodies a claim that is literally false. This tension can be resolved, fictionalists believe, by adopting the view that the claim does not intend to highlight a literal truth, but instead should be understood as a useful fiction.

While fictionalism and rationalizations appear to aim at resolving similar types of tension, the tension resolved by fictionalism is metaphysical while that of rationalizations is psychological. Rationalizations were never represented as aiming at literal truth, so the metaphysical tension present in the statement about the time the sun rose is absent from the rationalizations about the reasons the playable video game character acted as it did.

Rationalizations are unlike either of the approaches in philosophy described above: strategic inaccuracies in scientific models or fictionalism. Unlike fictionalism, rationalizations do not attempt to repair a metaphysical mismatch between an utterance meant to be taken at face value and the fact of the matter the utterance was meant to represent a truth about. Instead rationalizations are employed such that this tension is forever left unresolved. Unlike fictionalizations in scientific modeling, rationalizations do not attempt to highlight a deeper truth about the target system through the use of strategic inaccuracies. Rationalizations do not employ a mix of veridical and strategically non-veridical statements, they are entirely non-veridical. Rationalizations are thus entirely unmoored from the fact of the matter the natural language statements are supposedly about. Rationalizations do not make use of strategic inaccuracies or fictions in order to help individuals to come to recognize a greater truth about the explanandum. Instead, they serve to further conceal the truth behind natural language statements meant to have the appearance of an adequate explanation with none of its substance. These statements, strictly speaking, are fictional, but the purpose of the fiction is different from the purpose of those fictions employed in these

on earth at particular times of the day.

related philosophical contexts.

While there may be practical reasons why AI developers would find it appropriate to make use of rationalizations rather than genuine explanations, this does not imply that rationalizations have any value as scientific explanations. Rationalizations are an attempt to articulate "how possibly" explanations rather than "how actually" explanations. In the case of explanations of high-stakes automated decisions, "how actually" should be the standard, because anything less allows for the possibility of biased decision making that is masked behind a plausible and non-biased-sounding rationalization. Even if rationalizations were nearly always veridical articulations of relevant relations within the target system — they're not — the lingering possibility that they could be otherwise works against the stated purpose of building trust in automated decision making. The satisfaction of the need for an explanation of an AI's decision requires a "how actually" rather than a "how possibly" explanation. Rationalizations are not explanations.

1.3.3 Case Study Two: Attention Layers in Neural Networks

Attention mechanisms, introduced by Bahdanau et al.,[54] allow the training of a DNN in such a way as to focus the network's attention on specific input elements. In NLP tasks this may mean highlighting specific words in an input sentence that are key to decoding its meaning. Attention mechanisms can be incorporated into neural networks as another layer of the network as shown in figure 1.1. The weights of the attention layer are thought to correlate to measures of feature importance in the input: the input has some features that are more important than others (e.g. a key word or phrase), and if the attention layer is able to identify which features of the input are most important, this is thought to generate explanantia by discriminating between relevant and irrelevant inputs.

[54]Bahdanau, Cho, and Bengio, "Neural machine translation by jointly learning to align and translate".

Allowing the DNN to focus on the more important parts of the input increases the accuracy of the output. Attention mechanisms do a very good job of increasing the accuracy of NLP-based tasks that make use of them. Vaswani et al.[55] point this out to great effect in their appropriately named paper "Attention is all you Need." Networks made up almost entirely of attention mechanisms actually outperform other types of DNNs including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) that merely make use of a single attention layer. Due to the effectiveness of models that make use of attention, it is reasonable to ask if the feature importance identified by attention mechanisms amounts in some way to an explanation of the output of the model.

If attention were to be used as an explanation, the explanandum would be the output of the DNN, and the explanans would involve an appeal to the attention layer, which points to specific (and more relevant) input elements. The attention layer appears to be explanatory because it indicates which parts of the input were most important in the creation of the output. An answer to the question "why did the model make this classification?" could be simply, "these key words in the input string indicate it." For those evaluating these systems for explanatory value, this often appears to be a plausible explanation, though as I will discuss, there are good reasons for doubting that this is true.

1.3.4 Critical Responses from Computer Science

Jain and Wallace[56] argue that the output of the attention layer cannot serve as an explanation of the underlying DNN because it is possible to intentionally interfere with the way the weights of the attention layer are set (called "adversarial weighting") in such a way that the underlying DNN produces the same output as it did under non-adversarial weighting while the adversarial attention layer indicates the importance

[55] Vaswani et al., *Attention Is All You Need*.

[56] Jain and Wallace, "Attention is not Explanation".

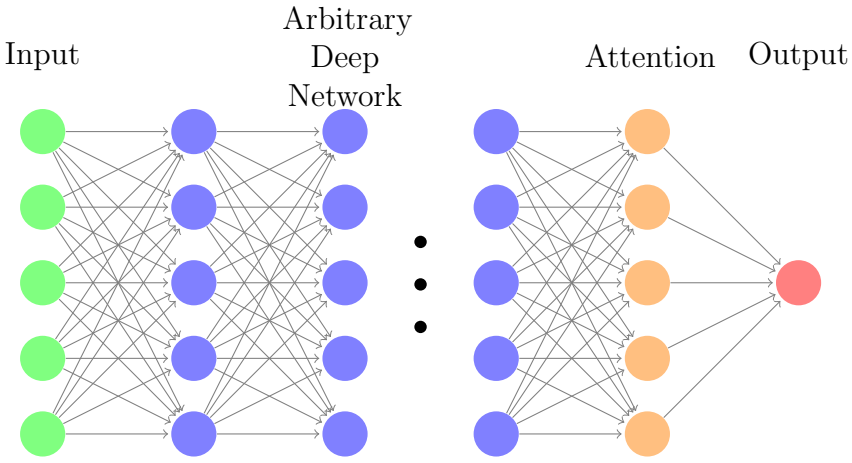

Figure 1.1: An arbitrary deep neural network with an attention layer. Researchers have used attention weights to generate explanations. Jain and Wallace ("Attention is not Explanation") scramble attention weights and show that output remains stable; Serrano and Smith ("Is Attention Interpretable?") omit highly-weighted attention nodes entirely while maintaining output stability

of entirely different — and obviously unimportant — elements of the input data. Jain and Wallace, in describing the motivation behind adversarial attention, explain that

> "the intuition is to explicitly seek out attention weights that differ as much as possible from the observed attention distribution and yet leave the prediction effectively unchanged. Such adversarial weights violate an intuitive property of explanations: shifting model attention to very different input features should yield corresponding changes in the output. Alternative attention distributions identified adversarially may then be viewed as equally plausible explanations for the same output."[57]

Jain and Wallace argue that if it is the case that attention weights are explanatory, counterfactual attention weight distributions which nevertheless produce the same model output should also be considered explanatory. This is a problem when the original and alternative explanations are inconsistent or contradictory. The existence of two contradictory, yet plausible explanations derived from attention weight

[57] Jain and Wallace, "Attention is not Explanation", p. 3548.

distributions which produce identical model output calls into question the very idea of an explanation derived from attention weight distributions.

Put another way, the overall DNN can be imagined as a set of propositions (D_1), and the attention layer as another set of propositions (A_1). In order for attention-derived explanations to work, one would expect that the output of the DNN, in this example a proposition P_1, should be consistent with D_1 and A_1. When the attention layer is adversarially weighted, it produces a new proposition A_2 which is consistent with both D_1 and P_1 but inconsistent with A_1. For any A_n that is consistent with D_1 and P_1, it is assumed that A_n is an explanation, but if A_1 is inconsistent with A_2, both cannot be explanations. The existence of these two inconsistent propositions that are still somehow consistent with D_1 and P_1 calls into question the use of either as an explanation of P_1.

An example discussed by Jain and Wallace is the use of a DNN to gauge whether a movie review is positive or negative. The DNN outputs a number between 0 and 1 with 0 being very negative and 1 being very positive. The attention layer indicates which words in the movie review (the input) are supposedly more important in determining this output. Under the non-adversarial case, a word like "waste" would be indicated as the most important, whereas under the adversarial weighting, a word like "was" would be indicated as the most important. In both the adversarial and non-adversarial cases, the network produced an identical score for the review.

While the attention weights were set adversarially, they still represent a configuration that could have occurred during the non-adversarial training of the network. That is to say that because the model produced identical output, the output in both cases is consistent with the training data — the model still correctly determines whether the review was negative or positive. It is possible, in training a neural model under normal conditions, that either the adversarial or non-adversarial attention layer weight distributions could have occurred. If one expects that the attention layer can

serve as an explanation of the overall model, it must be the result of the ability of the attention layer to identify the most important features of the input data, but if selectively modified attention weights can produce the same model output as the actual attention weights, it is difficult to see in what sense the attention layer could possibly generate an explanation. Jain and Wallace conclude that it cannot. Their paper is appropriately titled "Attention is not Explanation."

Serrano and Smith[58] make a similar argument, agreeing that attention is not explanation. Instead of assigning randomized weights to the attention nodes, the authors selectively deleted many of the highest weighted — that is the supposedly most important — attention nodes. Under these conditions the model still produced the same output. The experiment demonstrates that if adversarial attention weightings using data that should adversely affect the neural model's accuracy has no such effect, the ability of the attention layer to discriminate between important and unimportant inputs is called into question, and so must be any explanations that are derived from attention.

Both of these papers relied on counterfactual analyses of the attention layer in order to come to their conclusions: if the attention weights had been different in such and such a way, the attention layer would have identified a different set of input features, while the model's output would have remained unchanged. Implicitly, both are appealing to an interventionist account of explanation. They are attempting to determine the pattern of counterfactual dependence among the variables in the DNN. As I show below, due to the complexity and lack of interpretability of the systems to which this analysis is being applied, the use of the interventionist account here is inappropriate, and is not likely to lead to the development of XAI.

[58]Serrano and Smith, "Is Attention Interpretable?"

1.3.5 Why Attention is not Explanation

It is because of the implicit use of an interventionist account of explanation that previous attempts to use attention as explanation have failed. For reasons that will be explained in this section, surgical intervention on attention weights is impossible, which makes the use of causal explanation of DNNs problematic. Both Jain and Wallace and Serrano and Smith assume that all successful explanation must be causal. When the only acceptable explanations are causal, and when surgical intervention is impossible, explanation will be impossible. This is why attention is not explanation.

Bokulich defines 'causal imperialism' as the view that "all scientific explanations are causal explanations".[59] There appears to be a large amount of causal imperialism in XAI in general — most attempts at XAI make use of causal explanations exclusively, assuming that anything other than a causal explanation is fictional akin to the rationalizations described in section 1.3.1. The bar for explanation under these conditions is so high that some authors have advocated for abandoning the project of developing explainable models entirely, opting instead only for models that are interpretable.[60] There are simpler models that exist that are interpretable, such as decision trees, but they are generally less effective than more complex black box models. The tradeoff with these models is that a causal explanation can be more readily derived when a model is interpretable, because a pattern of counterfactual dependence within the model is easier to discover.

Given their complexity, a causal account of explanation that successfully explains DNNs is likely to be impossible because a pattern of counterfactual dependence cannot be located. The high number of nodes in a DNN, each with an associated weight, is not human parsable, and thus a complete account of causal relationships among nodes will also be non-parsable by humans. As just one example of how difficult it can

[59] Bokulich, "Searching for Non-Causal Explanations in a Sea of Causes", p. 141.

[60] Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead".

be for a human to parse the causal relationships within a large DNN, Google recently trained a neural model with over 600 billion parameters.[61] Simply iterating over that many parameters, to say nothing of understanding the causal relationships among them, would be an impossible undertaking even for a team of humans. The question of the meaning of interpretability is not easy to answer. Under some accounts, a network may be considered interpretable if the creation of a machine-generated account of the relationships between parameters is in principle possible. Under other accounts, interpretable implies interpretable *by people*. Under either of these views, however, the largest DNNs lack interpretability. AI that is non-interpretable will necessarily also be non-explainable under causal accounts, because to say that a system is non-interpretable is to say that a pattern of counterfactual dependence cannot be established for that system. This follows directly from the definition of non-interpretability. A non-interpretable system is a black box system; when the inner workings of a system are unknown, the causal relationships between that system's components cannot be established. Given the failure of causal accounts in the development of XAI, non-causal accounts of explanation should be explored instead.

The criticisms of attention as explanation from Jain & Wallace and Serrano & Smith implicitly make use of an interventionst account of causal explanation similar to that proposed by James Woodward.[62] Because the criticisms of attention as explanation attempt to establish the existence of empirically verifiable causal patterns that hold between the explanandum and those factors without which it would not have occurred, it fits within Woodward's framework. Woodward explains that "an intervention can be thought of as an idealized experimental manipulation which changes C 'surgically' in such a way that any change in E, should it occur, will occur only 'through' the change in C and not via some other route".[63]

[61] Lepikhin et al., *GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding*.

[62] Woodward, *Making Things Happen: A Theory of Causal Explanation*.

[63] Woodward, *Making Things Happen: A Theory of Causal Explanation*, p. 119.

In order to determine the existence of causal relationships between variables in a system of variables, the relevant variables are subject to manipulation. Successful explanations, on this account, require that targeted manipulations of relevant system components cause changes in the output of that system when the system output is the explanandum. If manipulations of these parts cause changes to the system's output, the core elements of an explanation are already present. Because the critics of attention as explanation were able to modify seemingly relevant variables without changing the system output, they concluded that deriving an explanation from attention is inappropriate.

The criticisms of attention as explanation implicitly appealed to a view similar to the interventionist account of explanation, but one without a requirement that some variables in the system be held invariant such that the interventions on the system are surgical. Following this requirement ensures that the explanation which is eventually generated can't be superseded by another more plausible explanation related to variables which were not controlled for. In the social sciences, for example, a study of the effects of diet on longevity that does not control for income is likely to be tainted by many spurious connections between variables that are better explained by the relationship between income and longevity than between diet and longevity. Without holding the extraneous variables invariant, the appropriate pattern of counterfactual dependence cannot be established. The absence of this requirement in the criticisms of attention as explanation may account for the results of these experiments: the discovery of nonsensical alternative explanations derived through the same means, which allowed the researchers to cast doubt on both sets of explanations. The situation does not improve significantly when surgical intervention is used; the problem with applying this approach to a DNN is that the number of interconnected nodes is so great that engaging in a surgical intervention on any one particular node is likely to be impossible as its value cannot be disentangled from the values of each other

node.

To be clear, it may be possible to write a program that works through each node of a DNN and tweaks the associated weight, noting the results. This process, however, would fail to generate a causal explanation for two reasons: first, it isn't enough to change the value of one node at a time; what must accompany this change in order for the manipulation to be considered surgical is for all other associated nodes that are also causally relevant to the end result to be held constant. Holding *all* other nodes in the network constant would not reveal the causal relevance of the target variable to the entire system (when other nodes are not held constant) due to the possibility that a change to the target node may cause a change in an intermediate node that then causes an important change in the output. If A and B both influence C, C influences D, and D influences E, then it makes no sense to hold A,C, D, and E constant while intervening on B. The purpose of the intervention on B is to track B's effects on C, D and E. This cannot be done if C, D, and E are held constant. In this system, the only variable that should be held constant is A. The problem with writing a program that iterates through each node and tweaks its weight is that it won't reveal the pattern of counterfactual dependence within the system because the causally relevant variables are unknown. A program which could reveal the pattern of counterfactual dependence would need to test every possible combination of variant and invariant nodes, substituting multiple values for each variant node in each iteration. Problems of computational complexity prohibit working through a system in such a brute force manner.

The second reason such a program would fail to generate a causal explanation is that any such explanation, even if it could programmatically generate the pattern of counterfactual dependence within the DNN, would likely be as complex or more complex than the DNN itself. Part of what a good causal explanation does is highlight the relevant elements of a system and describe how those elements fit together. The

hypothetical program which intervenes on each node of the network may manage to create an account of how *every* element of a system fits together (as I have argued above, I think this is unlikely) but it cannot extract only the relevant elements of the system to include in the explanation. Providing the entire frog genome in response to a question about why frogs croak is not a good explanation. It isn't enough to generate a machine-readable explanation; any meaningful explanation must be human understandable.

When making this limitation explicit — that surgical intervention on a single node of a DNN is not possible — the outcome is the same. It is still the case that attention is not explanation, but for a different reason. In this case attention is not explanation because under the interventionist framework, it is impossible to engage in surgical intervention on a DNN, and it is thus impossible to find a pattern of counterfactual dependence among the relevant variables within the DNN.

Under the manipulability account of causal explanation, surgical intervention is a method of testing counterfactual conditionals of the form, "if I were to change X in such and such a way, the result would be Y." Actually manipulating the value of X tests the truth of this conditional. Attention is only one part of a larger system of variables. The relevant system in this case is not attention alone, but attention in addition to the DNN itself. When surgical intervention is impossible, all counterfactuals are rendered unintelligible since surgical intervention is in one sense merely the testing of a counterfactual conditional. To say that surgical intervention on a given system is impossible is to say that the truth of counterfactual conditionals about that system cannot be tested, and thus the truth of those counterfactual conditionals is unknowable.

Of the two case studies explored in section 1.3.1 and section 1.3.3, what initially appeared to be the more plausible approach (the use of causal explanations through attention mechanisms in DNNs) now appears to be a dead end. While the use of ratio-

nalizations explored in section 1.3.1 has clear flaws, a factor motivating the approach, the desire to avoid the messy business of attempting to build causal explanations of DNNs, may have been correct. In the following section I will explore the possibility of applying non-causal explanations to DNNs.

1.3.6 Why Not a Partial Causal Explanation?

My position thus far has been that a full causal explanation is, for most neural nets in most cases, impossible in principle. An observer may wonder why it is necessary to have a *full* explanatory account that can trace *all* of the relevant causal relations through the entire network. After all, this type of complete causal account is not generally required for explanations of physical, chemical, biological, or human behavior. For instance, it is not necessary to have a complete understanding of the internal neural processes in the brain in order to explain why an individual voted for one candidate rather than another. In this instance, the production of a complete causal explanation does appear to be both unlikely and unnecessary. If one were to produce a fully uninterrupted account of all of the causally relevant factors leading up to that decision, there would be much to account for — certainly so much as to prohibit the possibility in principle of producing the explanation. Why then should we expect as much for causal explanations of deep neural networks?

This objection misses the mark for several reasons. First, if taken to its logical conclusion, it implies a belief in hard physical reductionism about the universe. If every causal factor must be accounted for before a successful causal explanation is possible, physics must be complete. But this is clearly not the case. So, given a charitable reading of the objection we ought not assume that this is necessary. Perhaps those who object on these grounds instead mean to say that in searching the *entire deep neural model* for causally relevant relationships before settling on an explanation, I am the one who is assuming physical reductionism.

I do not believe this to be the case either. I am not suggesting and have not suggested that one must account for every causal factor within the deep neural network between the input and the output, but instead only those that are causally relevant to the output. The problem with this type of explanation is that determining precisely which factors are relevant is the core of the problem, and when, as is the case within a deep neural network, not all of those factors are even known and are thus unavailable as candidates for consideration for causal relevance, we brush up against the possibility, as I have suggested elsewhere, that causal accounts of explanation are in-principle impossible in the case of many deep neural networks.

the target system in the example about explanations that answer the question of why a person voted for one candidate over another is not the brain, but myriad social and political factors in a social/political system. The target system in the question of the explanation of the neural net is the neural net. The decision to focus on the logical relationships in the neural net is no more folly than would be a focus on the political party of the voter in the human decision example.

No one is seeking explanations of the inner workings of the human brain and its causal relevance to the decision of which candidate to vote for, but we are, in fact, seeking explanations of the causal relevance of the structure of the neural network to the output of the neural network. To ask why we must demand such a robust low-level causal explanation in the case of the neural net but not in the case of human social choices is to miss the fact that the relevant question about the neural net is a question *about* a deep, low-level system, whereas the social question is not. Providing causal explanations in each case would thus demand different levels of specificity, and in the case of the relevant causal relations within the neural network, a very high degree of specificity is indeed required before a causal explanation will be successful. This is not so with causal explanations of simpler systems.

In fact, I am not suggesting that we should require, for instance, a full account

of the causal relations at the processor level, or at the level of the electricity flowing in and out of the processor, which would be analogous to asking questions about the brain in order to articulate a causal account about a human decision. When one asks a question such as "why did this person make this choice about this social system?" the relevant causal factors include personal preferences, social expectations and societal norms, political factors and so on. When one asks a question such as "why did this computer program produce this output given this input?" the only possible factors worthy of consideration must include those such as the logical pathways within the program that produced the output given the input, and in the case of the deep neural network, that includes the possibly trillions of parameters within the model. These factors are, in fact, the bare minimum that one must consider in order to form a coherent causal explanation. In exploring these relationships, and in concluding that determining the causally relevant relationships is in-principle impossible due to the inability to determine any pattern of counterfactual dependence among those variables, I am exploring the minimum number of potential causal factors necessary in order to produce a working causal explanation of the target system. To explore causal relations at a lower level such as the processor or physical level would be too much, just as it would be too much to explore low-level questions about the brain in an explanation of simple human social decision making. In short, this objection makes a category error.

1.4 Applying Non-Causal Accounts of Explanation to XAI

Both the causal and rationalization approaches to XAI have so far failed to yield good explanations of the decision process happening inside DNNs. The use of rationalizations was an attempt to build psychologically satisfying rather than veridical explanations. The attention example did appear to come closer to an acceptable conclusion. Even if the conclusion was that attention is not explanatory, the discovery of

this fact advances the discussion and sets up the possibility for the discovery of other causal explanations in the future. For reasons I discuss below, the use of non-causal explanations is more appropriate for XAI.

The counterfactual theory of explanation (CTE) has causal and non-causal variants. Computer scientists have previously used causal CTE in attempts to build XAI. See, for instance, Wachter et al.[64] and Sharma et al..[65] These approaches suffer from many of the same problems identified by computer scientists as discussed in Section 1.3.4 and by philosophers as discussed in section 1.3.5. Reutlinger[66] proposes a pluralist extension of the CTE which would allow for both causal and non-causal explanations under the CTE. If it is possible to use a non-causal variant of the CTE to explain DNNs, it might be possible to overcome the objections described in sections 1.3.4 and 1.3.5.

Jain and Wallace and Serrano and Smith are implicitly invoking a form of counterfactual explanation by attempting to answer what James Woodward has called a what-if-things-had-been-different question (W-question).[67] The problem with attention as explanation that these authors are highlighting stems from the fact that certain W-questions *are* answerable: we know that there are some models which produce identical overall output with obviously false attention output. The relevant W-question in this case has already been answered. If the model's attention layer had been different in the particular way prescribed by the authors, the model's output would remain unchanged. The W-question is answered, yet it is simultaneously the case that surgical intervention is impossible because it is not clear if the reason for the observed output is causally connected to the addition of the adversarial attention weights. There was an intervention, but it was not surgical. As Reutlinger

[64]Wachter, Mittelstadt, and Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR".

[65]Sharma, Henderson, and J. Ghosh, "CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models".

[66]Reutlinger, "Extending the Counterfactual Theory of Explanation".

[67]Woodward, *Making Things Happen: A Theory of Causal Explanation*, p. 191.

notes, Woodward acknowledges the potential for the existence of situations where W-questions can be answered, but interventions cannot be performed; in these situations a non-causal, rather than a causal explanation is available.[68,69] This appears to be exactly the type of case that Woodward is describing. For this reason, non-causal explanations hold greater promise for DNNs than causal explanations.

Mathematical explanation, another candidate category of non-causal explanation of AI, comes, according to Colyvan and McQueen,[70] in two varieties: intra-mathematical and extra-mathematical. Intra-mathematical explanation is "the explanation of one mathematical fact in terms of other mathematical facts," while extra-mathematical explanation is "the explanation of some physical phenomenon via appeal to mathematical facts".[71] Extra-mathematical explanation holds great promise for XAI because all DNNs are mathematical. One possible problem is that the relationship between the math used to build AI models and the world is more complicated than, e.g., the relationship between the mathematics used for graph theory when representing the bridges in the city of Königsburg as a graph and the actual city of Königsburg. If an AI classifier is putting images in categories, it can be described and explained in mathematical terms, but the relevant question we seem to want answered isn't about the math, but about the connection between the math and the world. The question of how an AI knows the difference between strawberries and bananas isn't a question limited to its internal mathematical operations because it is also appealing — even if implicitly — to the actual difference between strawberries and bananas. The Seven Bridges of Königsburg problem can be solved with graph theory, but the explanation is still recognizable as representing the actual city of Königsburg. The connection between mathematics and the world in this case is clear, but it is not clear in the case of extra-mathematical explanations of AI.

[68]Reutlinger, "Extending the Counterfactual Theory of Explanation", p. 80.
[69]Woodward, *Making Things Happen: A Theory of Causal Explanation*, p. 221.
[70]Mark Colyvan and McQueen, "Two Flavours of Mathematical Explanation".
[71]Mark Colyvan and McQueen, "Two Flavours of Mathematical Explanation", p. 232.

The potential for the use of models as explanations has been disccused by Bokulich,[72] Batterman and Rice,[73] Morrison,[74] and Potochnik[75] among others. Model explanations are an exciting possibility for DNNs because DNNs produce models which are used in decision and classification tasks. If models can serve as explanations, the explanation for DNNs could be found in the trained models (referred to as deep neural models). One major problem with this approach is that, with the types of explanatory models discussed in the philosophy of science literature, the explanatory model either itself requires no explanation, or can be explained much more simply than can the phenomenon being modeled. There is a problem if the explanation of the deep neural network can be found in the deep neural model, because the trained model is as difficult to understand as the DNN itself. This pushes the need for an explanation up one level, but does not eliminate it. The explanation of the trained model will suffer from the same problems as the explanation of the DNN. An explanation of the DNN that does not also explain the model (which is ultimately responsible for decision and classification tasks) is not enough. It isn't just the DNN which requires an explanation, but the DNN and the model it produces.

1.5 Conclusion

Because of the high stakes of AI-based decision and classification tasks, explanations of DNNs, deep neural models, and the decisions and classifications they produce are necessary. Computer scientists have attempted to develop explanations of these systems, but their efforts are inadequately grounded in theories of explanation. The study of scientific explanation by the philosophy of science is well suited to this task. Non-causal accounts appear to have greater potential to explain DNNs than causal accounts. Non-causal variants of the CTE, extra-mathematical explanations, and

[72]Bokulich, "How scientific models can explain".
[73]Batterman and C. C. Rice, "Minimal Model Explanations".
[74]Morrison, *Reconstructing Reality: Models, Mathematics, and Simulations*.
[75]A. Potochnik, *Idealization and the Aims of Science*.

model explanations all have potential to provide explanations of DNNs in the future, though more work needs to be done before this is possible. The persistent problems surrounding explanations of DNNs point to problems with existing accounts of scientific explanation and indicate the necessity for the extension of existing accounts of scientific explanation or the development of new accounts.

Chapter 2 Science, Values, and Artificial Intelligence

In Cormac McCarthy's Novel *No Country for Old Men*, a serial killer named Anton Chigurh flips a coin to determine whether he will murder his victims. When he confronts a character named Carla Jean, he flips a coin and tells her to call heads or tails. She initially refuses, but eventually calls heads. When Chigurh reveals that the coin is showing tails, Carla Jean tells him that he's trying to blame the outcome of the coin toss rather than accept responsibility for his own actions. Chigurh replies by appealing to the fairness of the coin toss, "It could have gone either way," he says. Carla Jean tells him again, "The coin didn't have no say. It was just you."[1] Very frequently, the character of the discourse around high-stakes algorithmic decision-making, particularly through the use of artificial intelligence (AI), resembles this interaction in McCarthy's novel. Algorithms are often presented, like the coin toss, as an objective way to make decisions, but algorithmic decision-making using AI is not objective. It is deeply social and political, and for those impacted by the outcomes, it is personal. One of the most important factors determining the objectivity of algorithmic decision-making, over and above its fairness and transparency, is the social, political, and historical context surrounding its use.

A growing literature[2] centered around algorithmic bias has attempted to address these concerns as well as answer questions about how bias works its way into decision-making algorithms, how it can be removed once it is located, and how future algorithms can be engineered with these issues in mind. Implicit assumptions are made by the creators of these algorithms about the nature of objectivity and the possibility

[1] McCarthy, *No Country for Old Men*, p. 258.

[2] see for instance S. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*; O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*; Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*; Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*

of engineering and deploying objective algorithms. These assumptions must be interrogated. Because so much of the philosophical problem space around algorithmic decision-making centers on the concept of objectivity, answers to questions about algorithmic bias can be elucidated through an exploration of the nature of objectivity. This chapter will synthesize work done by feminist philosophers of science, philosophers of technology, and AI ethicists in order to form conclusions about algorithmic bias.

Broadly speaking, the goal of my focus on the relationship between science, objectivity, and algorithmic decision-making is to demonstrate that current projects which endeavor to engage in automated forms of algorithmic decision-making are, like the coin toss in the above example, social rather than technological projects that propose technological rather than social solutions. It is imperative that they be treated as such. Donna Haraway coined the term "god trick" to describe the impossibility of, "seeing everything from nowhere."[3] I use the phrase to describe the active, intentional implementation of the View from Nowhere. To implement the god trick is to simultaneously take the standpoint of standpointlessness and to assume the possibility of acting without also being the object of action, what Andrew Feenberg describes as "the paradox of action."[4] The impulse to flock to the use of algorithms as a way of shielding people from the consequences of a decision is a misguided attempt to apply the "god trick" of the View from Nowhere through computer automation. The addition of the algorithm does not make the success of the god trick more plausible.

An acknowledgement of and respect for the situatedness of these technologies within a broader social, political, cultural, and historical context must be foregrounded in the ongoing discourse around AI and similar technologies. The articulation of the nature of this context and the actual relationships between the people,

[3]Haraway, "Situated knowledges: The science question in feminism and the privilege of partial perspective", p. 581.

[4]Andrew Feenberg, "Ten Paradoxes of Technology", p. 8.

technologies, and power structures within it is essential for understanding the current philosophical problems associated with algorithmic decision-making.

The argument of this chapter will take the following form: (1) There is a problematic pattern of the implicit acceptance of a standpointless, context-free notion of objectivity emerging in a subset of the AI science literature. (2) The number of AI science publications with corporate-affiliated authorship, especially among "Big Tech"[5] firms, is growing at an alarming pace. (3) A peculiarity of Silicon Valley culture is the funding of new technology companies based on narratives about a world-changing new technology, often embodying claims about the technological achievement of objectivity through the View from Nowhere. (4) The proliferation of the standpointless, context-free notion of objectivity in the AI science literature is a consequence of the combination of the growing corporate influence on AI science and the cultural incentives in Silicon Valley to adopt the View from Nowhere as the only possible view of objectivity. (5) Correcting this flawed notion of objectivity in the AI science requires, in some measure, addressing the source of the view within the corporate world. While rejecting the View from Nowhere and adopting feminist views of objectivity in this space would largely resolve the problem, doing so is not feasible because of the top-down, hierarchical power structure of corporations under capitalism. Meaningfully improving the objectivity of AI science can thus most readily be accomplished through the democratization of workplaces where this science is most frequently done. Worker unionization at Big Tech firms will substantially improve the objectivity of AI science.

[5]I use the term "Big Tech" to mean very large market capitalization ("mega-cap") technology companies. A mega-cap company typically has a market capitalization over $200 Billion. This includes companies such as Apple, Microsoft, Alphabet (Google), Amazon, NVIDIA, Meta (Facebook) and others.

2.1 False Objectivity and the View From Nowhere in AI

The recent surge in the popularity of AI, due mostly to the success of neural networks,[6] has brought with it a resurgence in the View from Nowhere and the value-free ideal. The idea of the View from Nowhere is that objectivity in the analysis of something is possible without reference to one's historical situatedness, reliance on that history, or subjective experience within that historical context. Robert Scharff has argued that it is impoosible to be"an unhistorical or extra-historical thinker."[7] Scharff argues that the View from Nowhere is descended from Descartes' epistemological turn, and that the main concern with the Cartesian legacy is the persistence in philosophy of the standpoint upon which the Cartesian project was founded.[8] The problematic standpoint being one which presupposes the possibility of being "standpointless or presuppositionless,"[9] which nevertheless remains a standpoint. Scharff, citing Heidegger, explains that "there really is a View from Nowhere; it is just not a view from nowhere... the very idea of a position that achieves 'freedom from all standpoints... is itself something historical, something bound up with Dasein... and not a chimerical in-itself outside of time.'"[10] In short, the View from Nowhere is the standpoint that it is possible to engage with the world without having to do so from any particular standpoint. A related concept in the analytic philosophy of science is the value-free ideal. This is the notion that "social, ethical, and political values should have no influence over the reasoning of scientists, and that scientists should proceed in their work with as little concern as possible for such values."[11] The value-free ideal, like

[6] Sejnowski, *The Deep Learning Revolution*, p. 171.

[7] Scharff, *How History Matters to Philosophy: Reconsidering Philosophy's Past After Positivism*, p. 2.

[8] Scharff, *How History Matters to Philosophy: Reconsidering Philosophy's Past After Positivism*, p. 2.

[9] Scharff, *How History Matters to Philosophy: Reconsidering Philosophy's Past After Positivism*, p. 2.

[10] Heidegger and Buren, *Ontology—The Hermeneutics of Facticity*, p. 64; as cited in Scharff, *Heidegger Becoming Phenomenological: Interpreting Husserl through Dilthey, 1916–1925*, p. 40.

[11] Douglas, *Science, Policy, and the Value-Free Ideal*, p. 1.

the View from Nowhere, is not achievable, and should not be a desideratum of any productive scientific research program.

The current discourse around bias in AI reveals that attitudes about the objectivity of algorithms are deeply intertwined with implicit beliefs about the utility of both the View from Nowhere and the value-free ideal. Recent AI applications have perpetuated the illusion that bias can be eliminated from science through the use of purportedly "objective" algorithms;[12] if we remove the humans from the decision-making process, the thinking goes, then we remove the bias.[13] In reality, neither is possible. Doing so would be to assume the standpoint of standpointlessness. The fallacy is to imagine that even if humans cannot be objective, the algorithms we create can be, and AI can thus assume the View from Nowhere on our behalf. To believe that doing so is appropriate or even desirable is to perpetuate the value-free ideal. Of course none of this is possible; humans create and implement the algorithms, and they — consciously or unconsciously — incorporate their biases into them. Philosophers have rejected the View from Nowhere and the value-free ideal before, and in light of these attitudes about AI, they must be rejected again in this new context. It is important to recognize that AI is not a free lunch — it cannot remove a scientist's bias, it cannot disentangle science and values, and it cannot create pure objectivity ex nihilo. The people who create AI are starting from a standpoint, and so is the AI that they deploy.

One particularly powerful example of the problem at hand — the attempt to achieve objectivity in decision-making by foisting the responsibility for those decisions onto the purportedly objective algorithms that make them — can be found in a recent wave of "neo-physiognomy"[14] AI papers. These papers have attempted to, among other things, determine an individual's sexual orientation from the way

[12]Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead".

[13]Khademi and Honavar, "Algorithmic Bias in Recidivism Prediction: A Causal Perspective".

[14]Blaise Agüera y Arcas and Todorov, *Physiognomy's New Clothes*.

they walk,[15] determine a child's emotional state through the use of facial recognition software,[16] determine a person's gender based on an image of their face,[17] determine a job applicant's suitability by applying NLP to their resume,[18] determine an individual's trustworthiness based on their facial features,[19] or determine the likelihood that an individual will commit a crime based on a photograph of their face.[20] These ideas replicate racist and sexist practices from the history of science,[21] obscured by technical jargon. They often rely upon seemingly commonsense assumptions, such as the assumption that a person's inner mental state can be determined based on their facial expressions, for which there is minimal or no evidence.[22]

These applications perpetuate existing discriminatory attitudes and practices both within and outside the sciences, and bear a strong resemblance to similar abuses in the history of science, e.g. scientific racism.[23] All rest upon the assumption that the biased application to which they are put is somehow canceled out by the fact that the decisions are made by an algorithm rather than a human. While it would be wildly inappropriate for a person to walk down the street and guess at each passing stranger's sexual orientation, the clearly social character of such an exercise, for some reason, seems to disappear into the background when rendered in a programming language.

There is an underlying metaphysical assumption behind the creation of these types of algorithms that there genuinely is a fact of the matter to discover, and that this

[15]Wang and Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images."

[16]Marechal et al., "Survey on AI-Based Multimodal Methods for Emotion Detection".

[17]Akbulut, Şengür, and Ekici, "Gender recognition from face images with deep learning"; Ng, Tay, and Goi, *Vision-based Human Gender Recognition: A Survey*.

[18]Van Esch, Black, and Ferolie, "Marketing AI recruitment: The next phase in job application and selection".

[19]Safra et al., "Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings".

[20]Wu and Zhang, *Responses to Critiques on Machine Learning of Criminality Perceptions (Addendum of arXiv:1611.04135)*.

[21]Harvard University Library, *Scientific Racism*.

[22]Crawford, *The Atlas of AI*, p. 171.

[23]Harvard University Library, *Scientific Racism*.

fact exists independently of the perspective of any particular human. There is an associated underlying epistemological assumption that this fact of the matter is both discoverable and knowable.

Another example of an AI project which relies heavily upon these types of assumptions is an algorithm released publicly in 2020 called PULSE (Photo Upsampling via Latent Space Exploration).[24] The project was meant to take low-resolution images as input and produce high-resolution images as output, similar to the "zoom-enhance" style of image upsampling featured in many works of science fiction.[25]

Very shortly after posting the pre-print of their article, Menon et al. received swift criticism from data scientists on social media. The critics pointed out that using human-recognizable, low-resolution images of people of color as input to PULSE very frequently output upscaled images of white people. A viral example posted to Twitter showed a low resolution image of President Obama upsampled into an image of a white man.[26] Data scientist Robert Osazuwa Ness was able to reproduce the findings, and additionally demonstrated the same problem with low resolution images both of himself and Congresswoman Alexandria Ocasio-Cortez.[27]

Given the above examples, *that* the PULSE algorithm shows troubling signs of racial bias should not require further argument, but *how* the bias came to be a part of the algorithm has been a source of debate among data scientists, machine learning practitioners, and philosophers of science. The immediate response to these criticisms was an appeal to biased training data,[28]. The bias crept in to the finished model,

[24]Menon et al., *PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models*.

[25]This is a trope commonly made use of in the television show CSI: Crime Scene Investigation. Scenes generally involved a detective providing a low resolution image to a technician who used a computer to recover impossible amounts of detail from it by, e.g. zooming in to a small reflection in a window on a grainy image to read a car's license plate.

[26]Vincent, *What a machine learning tool that turns Obama white can (and can't) tell us about AI bias*.

[27]Vincent, *What a machine learning tool that turns Obama white can (and can't) tell us about AI bias*.

[28]In "supervised" machine learning (ML) researchers present a computer program with labeled examples of the types of data it will work with in production. The program finds statistical simi-

they argued, because the model was trained with more pictures of white people than pictures of people of color. Because of the composition of the training data, the model is biased in favor of outputting upscaled images of white people.

As professor of information studies Safiya Noble has argued, data is a social construction.[29] There is no unbiased data because all data must be collected by a human (or a tool built by a human), and what is determined as worthy or unworthy of inclusion in a data set is already a social and political decision. For example, primatology research, prior to the 1960s, was conducted primarily by men, who often focused the attention of their data gathering on the social behavior of male primates. When significant numbers of women began working in the field in the 1970s, the data collected suddenly changed to include interesting and important observations about female primates.[30] If there is a fact of the matter to discover in the data, which facts are discoverable depends quite heavily on what data was and was not collected, since it is not possible to collect data with perfect fidelity. But this is, in any case, largely irrelevant because the notion that there is a fact of the matter to discover in the first place is questionable; a discovery requires a discoverer. Facts cannot reveal themselves any more than science can conduct itself. Science and Technology Studies scholars have argued for years that science and society are in a state of co-production.[31]

In response to rising concerns about bias in AI, some computer scientists have advocated for simple solutions to the problem of AI bias like sanitizing and de-biasing

larities and differences between labeled inputs and outputs and produces a "trained" model. This labeled data used to produce the model is referred to as "training data." For instance, in order to distinguish from photos of individuals with and without hats, the program may analyze 10,000 labeled photos each of people with and without hats. Using this training data, a model is produced which is capable of determining if a person in a new (unlabeled) photo is wearing a hat. Another common training method is "unsupervised" machine learning, also called "self-supervised" machine learning. In this case the training data is presented without labels. In cases both of unsupervised and self-supervised learning, the researchers training the algorithm have desiderata for the trained model that influence the training process.

[29] S. U. Noble, "Your Robot Isn't Neutral", p. 205.

[30] Haraway, *Primate Visions: Gender, Race, and Nature in the World of Modern Science*, pp. 286–288.

[31] Jasanoff, *States of Knowledge: The Co-Production of Science and the Social Order*, p. 17.

training data.[32] This is an attempt to address the problems described above associated with biased data collection practices and the resulting biased data sets. While debiasing training data is a good start, this alone will not solve the problem of the biased output of trained machine learning models, because training data isn't the only part of the system where bias can exist. It is possible for the training data to be scrubbed of all mentions of, e.g., income while the trained model remains biased in favor of the wealthy when applied in real-world scenarios. This problem pre-dates machine learning. Researchers have been sounding the alarm for decades about the issue of proxy variables in aptitude testing such as the well-known correlation between family income and SAT scores.[33] The problem has little to do with what sensitive data is collected and much more to do with answering the question of why higher family income is correlated with better test performance.

The factors leading to the introduction of bias are much more broad than data collection practices — bias exists at every level, but ultimately it is a consequence of the inequalities present in the social world, which, intentionally or unintentionally, get built-in to our technologies. This is not to say that better training data shouldn't be a priority, but rather that the solution to AI bias cannot lie solely in the training data. The solution to the big-picture problem requires that researchers who wish to develop more objective algorithms spend more time engaging with existing work in the philosophy of science and the philosophy of technology centered on the nature of scientific objectivity and technological neutrality.

No technology is neutral, and this includes trained machine learning models, with or without sanitized training data. This technology, like all technology, is always already political, and cannot be made otherwise. By this I mean that technology presents itself to subjects, not as a neutral non-social object, but in the form of the potential social change embodied in and enabled by the object. An object which does

[32]Chen et al., "AutoDebias: Learning to Debias for Recommendation".
[33]Zwick, "Is the SAT a 'Wealth Test'?"

not manifest this way is not technology, hence technology is always already political.

The AI bias problem is bigger than just unintentional AI-enabled racism, sexism, homophobia, or transphobia. It represents an epistemological crisis that is related to an ongoing social justice crisis. There is a clear relationship between science in practice and human values. What the scientific community accepts as legitimate — and stamps as legitimate through publication — is already part of a complex fabric of social relations,[34] and it is important to articulate and examine the nature of the relationship between scientific practice and social justice. There appears to be a re-emergence of an old problem in these new AI papers: the illusion that the use of algorithmic solutions to problems in science renders science value-free, and that the algorithms are themselves neutral and not subject to human biases.

Bias is not merely an emergent property of complex systems. It is not a mere nuisance or a coding error that can be commented out or patched in later versions. Bias exists in our algorithms because people put it there, even if unconsciously. This is a consequence of what Donna Haraway describes as the second of three core boundary breakdowns, the boundary between the human-animal and the machine[35] The bias in the code is an extension of the bias of people, and it is impossible to genuinely distinguish between the two. The former cannot be removed without solving the latter, and human bias is ineliminable. Algorithmic bias cannot be wished away through superior technology, more efficient coding, or sanitized training data. The separation of human bias from algorithmic bias is not and has never been possible because algorithmic decisions have always been human decisions. To argue otherwise is to employ the god trick[36] of the View from Nowhere: to assume that the system upon which humans exert their technological influence is separate from the system

[34]Laudan, *Science and Values: The Aims of Science and Their Role in Scientific Debate*; Longino, *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*.

[35]Haraway, "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century", p. 11.

[36]Haraway, "Situated knowledges: The science question in feminism and the privilege of partial perspective", p. 581.

within which they live. As Andrew Feenberg notes, "when we act technically on an object there seems to be very little feedback to us, certainly nothing proportionate to our impact on the object. But this is an illusion, the illusion of technique."[37]

Every human use of technology is and always has been political, because all technological interventions have built-in feedbacks that impact the broader system beyond just the user of the technology and the target of the intervention. Feenberg argues that the purpose of the human use of technology is to influence the target more than it influences us in order to avoid the Newtonian "equal and opposite reaction" of our interventions, but such avoidance is possible for only a finite amount of time.[38] Eventually the externalities[39] of our interventions accumulate and begin to exert their influence back onto us. This is what Feenberg describes as "the paradox of action."[40] The algorithmic decisions and the algorithms themselves emerge from and make their consequences felt from within the same system. There is nowhere outside of this system where the user of a technology can be permanently sheltered from the effects of its use. Much of the problem of algorithmic bias stems from the ease with which this fact is forgotten, and the extent to which it has been intentionally covered over.

As the opening example from *No Country For Old Men* should make clear, the type of technology — and more importantly, the level of complexity of that technology — used to facilitate decision-making has no impact on who shoulders the responsibility for the consequences of those decisions. The stakes involved in the outcome, from high-stakes life-or-death decisions to low-stakes is-this-a-picture-of-a-fruit-or-a-vegetable[41] type decisions, don't matter either with regard to the assignment of re-

[37] Andrew Feenberg, "Ten Paradoxes of Technology", p. 8.

[38] Andrew Feenberg, *Technosystem: The Social Life of Reason*, p. 3.

[39] in economics, an "externality" is a cost that is not borne by the producer of a good, but is instead a shared cost paid by a broader community. An example is air pollution, which is a cost of the production of some consumer goods, but the costs of air pollution are not paid directly by the producers of those goods. Instead they are paid by, e.g., the people who suffer from resulting chronic health conditions, the healthcare system which must treat them, and health and life insurance companies.

[40] Andrew Feenberg, "Ten Paradoxes of Technology", p. 8.

[41] Many applications of computer vision involve classifying objects as belonging to one category or

sponsibility for the consequences of the outcome. In all possible combinations within this matrix of high-tech to low-tech, high-stakes to low-stakes, humans are in charge of making and enforcing the decisions. The problem is how easily this fact is obscured, and the degree to which such obfuscation is made possible.

There are several good reasons to focus on the human as the moral agent in algorithmic decision-making rather than the algorithm. First and foremost is that moral agency is reserved for human beings and human beings alone, and is not something that is or should be ascribed to computational systems.[42] There are no situations where the moral responsibility for an action lies with any tool used to facilitate the action rather than the person using the tool.[43] There are no variations of the trolley problem that question the moral responsibility of the lever rather than the one who pulls it.

A possible objection states that by reserving moral agency for humans and denying the possibility of the moral agency of technologies, I am appealing to the neutrality of technology in the same way that gun rights advocates do when they say that "guns don't kill people, people do." The issue of moral agency for technologically mediated interventions is distinct from the issue of the neutrality of technology. As I have argued, technology is not and cannot be neutral, but this can be the case without ascribing moral agency to technology. In Langdon Winner's famous example of an artifact with politics, the low overpasses on the parkways on Long Island which allow cars to pass but not busses,[44] it is simultaneously the case that the overpasses have

another. One important use of this technology is in autonomous vehicles which must be able to tell the difference between stop and yield signs, vehicles and pedestrians, same-direction and opposing traffic etc. Other uses of this technology carry much lower stakes, such as determining whether or not someone's pet is in a picture, or whether an image is of a fruit or a vegetable.

[42]Noorman, "Computing and Moral Responsibility".

[43]There are situations where a tool is used appropriately but still causes harm to the user, for example, cars sold with faulty brakes, or cribs that have a tendency to tip over. In these instances, responsibility for the failure of the technology lies with neither the user nor the technology itself, but with someone on the design or development team that created, quality-checked, or sold the product. In any case, even when technologies behave unexpectedly, they have no moral agency.

[44]Winner, "Do Artifacts Have Politics?"

politics and that they lack moral agency. The moral agent responsible for the politics of the overpasses is their designer, Robert Moses. It would, of course, be absurd to blame the overpass itself rather than the person who designed it, but blaming the person does not render the technology neutral.

That the issue of the moral agency of complex AI decision-making systems appears to be an open question is a result primarily of the complexity of the technology, and the difficulty of understanding it. There was never a question about the moral agency of the coin toss, but if there is a question about the moral agency of a trillion parameter deep neural network, it is only in light of the technology's impenetrability. Despite the difficulty of understanding or explaining[45] the decision-making process of the most complex AIs, the AI's status as a moral agent is not in question. Philosopher of technology Mark Coeckelbergh explains that "an AI can take actions and make decisions that have ethical consequences, but is not aware of what it does and not capable of moral thought and hence cannot be held morally responsible for what it does. Machines can be agents but not *moral* agents."[46] Far from being a reason to assign moral agency to algorithms, the black box at the heart of many decision-making AIs should actually push us further in support of human decision-makers as the only moral agents possible in this complex system.

As the complexity of the technology increases and the capacity of the users of the technology to thoroughly understand it decreases, the weight of the moral responsibility on the user increases. This is the case because the user is not only deploying a technology that has unknown social effects, but the fact of the existence of those unknown effects is known. Since the inside of the black box is a 'known unknown,' its use is irresponsible. Simply put, there is no way for a human who uses algorithmic decision-making to deny their status as a moral agent and instead shift the moral responsibility onto the technology itself. Humans bear the ultimate responsibility of

[45]for a longer discussion of this point, see Chapter 1 of this dissertation.
[46]Coeckelbergh, *AI Ethics*, p. 111.

all algorithmically derived decisions.

While the novelty of AI makes this problem appear to be new, the existing literature in the philosophy of science and science and values has addressed very similar problems over the past 30 years. There is a tendency to ascribe a kind of privileged epistemological position to AI, as if it has access to knowledge that humans do not, or as if it can be asked a research question and be expected to produce capital-T Truth as output. But this is fantasy; AI can do no such thing. It suffers from the same limitations that all sciences suffer from, namely that science is a human practice, and is not immune to bias. The value-free ideal must be rejected all over again. Values inform scientific practice, and this is, when appropriately acknowledged, not a problem for science or scientists. The problem emerges when values inform scientific practice while at the same time those engaged in the practice come to believe that their actions are value-free and objective: the View from Nowhere. This is not possible in science, nor is it possible in AI. Values are a part of the development, use, and interpretation of AI.

At the core of the solution to the problem of AI bias is the explicit recognition that all technology is political, that the research and design process in the development of new technology is political, and that choices about how (and upon whom) to deploy technology are political. Real progress on the problem of AI bias cannot be made without widespread recognition of the interrelatedness of the social and political world with the world of science and technology. Computer scientists who work to develop new algorithmic solutions to existing problems are doing work that is entirely political. There can be no distinction between the social-political and the techno-scientific.

2.2 Corporate Affiliated Authorship in Computer Science

Another emerging problem in AI ethics, which is actually a subspecies of the type of bias described above, is the overlap between corporate-funded AI research and the

identification and articulation of ethical problems in AI. A large amount of research in AI is done, as in other disciplines, in the academy. Scientific advancements in AI research are also made by corporate actors in the tech sector, notably Google, Microsoft, Meta (formerly Facebook), NVIDIA, and others. A recent analysis of corporate authorship in the most highly cited AI papers shows both a high degree of corporate involvement in AI science and rapid growth in this involvement. From 2009–2019, corporate-affiliated authorship of the most highly cited AI papers grew nearly 3-fold from 24% to 71%.[47] Moreover, the share of "Big Tech"[48] involvement among the corporate-affiliated authors grew by nearly 400% in the same timeframe.[49]

The participation of corporate actors in science is not new; it has been common in chemistry, medicine, agriculture, and other fields for decades.[50] There is no question that contributions to science from corporations can be good. Examples include contributions from the invention of the transistor[51] to the development of the COVID-19 vaccine.[52] But there are legitimate concerns about corporate researchers being, at times, at odds with the aims of science. Transparency and information-sharing are assumed to be aims of science, and skirting these expectations through p-hacking or selective reporting have been widely condemned within the scientific community. Corporate actors involved in science, however, have what economists call "perverse incentives" to engage in these same types of behaviors, at odds with the aims of science, as a result of their many conflicts of interest. Corporations engaged in scientific research have been known to intentionally mislead the public, hide the harms of their practices, and resist regulation that would prevent such behavior.[53] The existence

[47] Birhane et al., *The Values Encoded in Machine Learning Research*, p. 8.

[48] defined as "large tech firms, such as Google and Microsoft"(Birhane et al., *The Values Encoded in Machine Learning Research*, p. 8)

[49] Birhane et al., *The Values Encoded in Machine Learning Research*, p. 8.

[50] Pithan, *Corporate Research Laboratories and the History of Innovation*, p. 1.

[51] Gertner, *The Idea Factory: Bell Labs and the Great Age of American Innovation*, p. 98.

[52] Pfizer, *Our Path to Developing the Pfizer-BioNTech COVID-19 Vaccine*.

[53] Legg, Hatchard, and Gilmore, "The Science for Profit Model—How and why corporations influence science and the use of science in policy and practice", p. 2.

of this tension between corporate profits and transparency in science is a serious problem.

Similar conflicts of interest in the recent history of science include the actions of fossil fuel companies researching climate change in the 1970s,[54] automobile manufacturing companies researching (or selectively refusing to research) automotive safety in the 1960s,[55] and tobacco companies researching the health and safety of cigarettes in the 1950s.[56] In each case there is a major, unambiguous conflict of interest, and the existence of this conflict raises questions of propriety, transparency, and research accuracy. The same is true with corporate-sponsored AI research, so in many ways this is an old problem, and not particularly surprising. AI is different in several key ways, however.

First, achieving state-of-the-art results in natural language processing (NLP) is extremely resource intensive, and when it comes to the large language models (LLMs) that currently dominate the state of the art in NLP, the amount of computing power necessary to develop, train, and run the models is often so large as to preclude the involvement of non-government researchers who don't have either an incredibly large research budget or a pre-existing financial interest in advancing the state of the art, as would companies such as Google. Tobacco and automobile companies didn't have a de facto monopoly on researching their products, but in many important ways, Google does. The concern is that the relationship between the need for computing resources and the ability to advance the state of the art in NLP means that only those who already have a financial and commercial interest in NLP will have a say in the research direction for the field as a whole. This is a challenge not only for science, but for democracy.[57]

[54]Hall, *Exxon Knew about Climate Change almost 40 years ago*.

[55]Nader, *Unsafe at Any Speed: The designed-in dangers of the American automobile*, pp. 58–59.

[56]Bero, "Tobacco industry manipulation of research.", p. 200.

[57]Ahmed and Wahed, *The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research*, p. 2.

The second way that corporate involvement in AI research is different from tobacco or automotive safety research has to do with the way AI research is conducted at universities. While it is true that not all work in this space is conducted in corporate labs, the research undertaken in academic settings is often conducted using tools developed by corporate AI labs. Even when universities do have substantial resources which give them access to the computing power necessary to create and train large models, it is unlikely that all of their work is taking place without coming into contact with or making use of corporate-funded AI research. An example of such an LLM is Google's Bidirectional Encoder Representations from Transformers (BERT),[58] created and open-sourced by Google in 2018, which has since served as the basis for significant additional academic research in NLP.[59] Especially given the fact that the largest, most successful, and most accessible models have been trained by Google and similar companies, attempting to develop a new model from scratch feels somewhat like reinventing the wheel. Academic labs will often default to using the tools that are already available, and in the case of AI that frequently means interacting with corporate-funded research.

The main concern about this structure is that if the ideas coming out of the academy take a back seat to those coming out of corporate labs, it creates a hierarchical power structure within science, and importantly, this structure is dictated by the flow of capital. The ideas developed at the university depend upon the ideas and tools developed by the corporation. But if this is true, and it is also true that there is often a form of self-censorship at the corporate level around ideas that threaten the potential for profit in the future, then there is a filter installed within the machinery of science that will not let any ideas that threaten the status quo through. This is a good thing for the corporation because it ensures that no science that threat-

[58]Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".

[59]there have been over 800 computer science pre-prints containing "BERT" in the title posted to arXiv.org since BERT was released by Google in November, 2018.

ens their revenue stream will gain traction. The concern, however, is that as checks against corporate power continue to erode in the United States and elsewhere, there is a chance that the share of power and control wielded by corporate science over academic science will continue to grow, endangering the continuation of the entire enterprise.

AI scientists, generally speaking, care about fairness, accountability, and transparency in their work[60], but when pushing their research in this direction, they often meet resistance from entrenched systems of power. This is why some AI researchers are calling for an analysis of power and control within AI research rather than a shift towards more research on fairness and accountability.[61] The problem with simply increasing the focus on fairness, accountability, and transparency under the existing power structure is that if power shapes research, then it also shapes research on fairness, which diminishes the potential for radical change rather than fostering it.

Answering the question of why algorithms are biased is not as simple as saying that the training data was incomplete, a frequent refrain heard in response to the types of bias discussed in section 2.1. Bias does not emerge in a vacuum; it is not a rounding error. It emerges within the context of the hierarchical power structures of both corporations and the academy. It emerges from a complex social system that includes ongoing problems with racism, sexism, and regular violations of worker rights. The defining feature of capitalism is its ability to facilitate the concentration of wealth — and thus power — in the hands of the very few. This concentration of power has unexpected feedback effects on systems that many would assume to be unconnected, including science. The machinery of power and control wielded by the corporate leaders in AI science can itself be seen as a technology, and as Feenberg's "paradox of action" states, we must always eventually feel the full impact of the use

[60]There are multiple, well attended, Association for Computing Machinery (ACM) conferences devoted to these topics, most notably Artificial Intelligence, Ethics, and Society (AIES) and Fairness, Accountability, and Transparancy (FAccT).

[61]Kalluri, "Don't ask if artificial intelligence is good or fair, ask how it shifts power".

of our technology.[62] The feedbacks of the use of the machinery of power and control by Big Tech companies are now being felt as disruptions in the flow of knowledge production through AI-science. These disruptions manifest as anomalous biases in our production AI systems, but these are no mere anomalies — their emergence was unavoidable as they exist at the intersection of Haraway's blurry interchange between the animal-human and machine.[63] The bias is not in the machine because there is no longer a distinction between it and us.

Bias is an irreducible consequence of human subjectivity, and it will always be a part of science. Its presence is, however, amplified by power imbalances in a knowledge production mechanism built and maintained largely by corporations beholden to capital. Without a thorough analysis of power structures within science, and an accompanying analysis of how that power is distributed both inside and outside of the academy, it will not be possible to solve[64] the problem of bias in AI. Critiques of AI and other technologies must be engaged in, not only from the standpoint of science and values in the philosophy of science, but through the lens of more explicitly social and political philosophy. This effort is particularly important within computer science and AI because, as historian of technology Mar Hicks observes, "the entire history of electronic computing is, as is the case with many technologies, intertwined with efforts at domination."[65] The existing power structures within science and technology must be acknowledged and addressed if the harms of AI bias are to be meaningfully reduced.

Plainly put, in some measure science is, as all human activities are, concerned with how to share and distribute social power.[66] The current balance of power in society

[62]Andrew Feenberg, *Technosystem: The Social Life of Reason*, p. 3.

[63]Haraway, "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century".

[64]the existence of a solution to the problem of bias does not imply the necessity of eliminating bias. As I have argued, I do not believe that is possible.

[65]Mar Hicks, "When Did the Fire Start?", p. 17.

[66]Bruno Latour famously said, echoing Carl von Clausewitz, that "science is politics by other means"(Latour, Sheridan, and Law, *The Pasteurization of France*, p. 229)

is inappropriately skewed in favor of a very select group of people, and many of the AI bias problems discussed in section 2.1 stem from this fact about our society and our politics. As one glaringly obvious example of the power imbalance in this space, consider the many contributions of women and people of color to science, technology, engineering, and mathematics (STEM) documented thoroughly, among other histories, by Mar Hicks[67] and Charlton McIlwain.[68] The contributions of these groups to STEM fields are undeniable, yet the celebrated figures in these disciplines are overwhelmingly cis white men. The issue is not a lack of participation by marginalized groups in STEM, but the structure of power within these fields in particular and society generally. This state of affairs is made worse as a result of multiple factors, including capitalism, which has a tendency toward the accumulation of power in the hands of relatively few. Bias and inequality are problems in AI because they thoroughly permeate society. It is not possible to address the former without addressing the latter.

Science is not disconnected from politics. If it is appropriate for scientists to oppose algorithmic bias, or to stand against racism and sexism, it is appropriate for scientists to loudly criticize corporate involvement in scientific research. This is a necessary step in order to begin to solve problems associated with bias in AI, and philosophers of science should advocate for this position without worrying that it will appear too political. Capitalism makes racism and sexism worse by facilitating the further concentration of power in the hands of the few. These same problems, despite our best efforts, are ultimately replicated in the lab. Any lingering reluctance to acknowledge as much is related to a continued attempt to adopt an entirely objective view of the world - the failed View from Nowhere.

The best recent example of the serious nature of the problems emerging from

[67]M. Hicks, *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing*.

[68]McIlwain, *Black Software: The Internet & Racial Justice, from the AfroNet to Black Lives Matter*.

the overlap between corporate power and scientific research in AI can be found in the sudden, late-2020 firing of Google's AI ethics team lead[69], Dr Timnit Gebru.[70] Several members of the team had co-authored a paper[71] about the environmental harms caused by large language models (LLMs) used in NLP by Google and others. As a result of the paper's publication, two authors were fired from the Google AI ethics team, apparently because the paper's conclusions were in direct contradiction to the interests of the authors' employer, Google. This incident demonstrates a large area of overlap in multiple ongoing philosophical problems including the role of scientific research on climate change, capitalism's influence on science, and the (often uncomfortably) close relationship between scientific and economic productivity.

The LLMs in use by Google which were the subject of the criticism leveled by Bender et al.[72] can themselves be seen as an attempt to sell the possibility of the god trick to a wide audience. Google wants its customers and the broader public to believe that its LLMs can see everything from nowhere, and to act without being the object of action. The importance of the god trick to Google's business model becomes apparent when one recognizes that the mere acknowledgement of the existence of externalities (like carbon emissions and energy consumption) associated with LLMs is what directly spurred Google to fire its AI ethics team. The work of Dr Gebru helped to situate LLMs within a broader social, political, and environmental context. When the public is able to see LLMs in this light, Google is prevented from executing the god trick. LLMs can no longer be presented as being capable of seeing everything from nowhere and acting without being the object of action. The publication of the Stochastic Parrots paper pulled back the curtain to allow the public to see the inter-relatedness of AI and the social and political world. Once LLMs are understood

[69]the rest of the team was later also fired or otherwise forced out.

[70]Metz and Wakabayashi, *Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.*; Tiku, *Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it.*

[71]Bender et al., "On the dangers of stochastic parrots: Can language models be too big".

[72]Bender et al., "On the dangers of stochastic parrots: Can language models be too big".

as being situated within this context, the god trick is more easily recognized as impossible. Because the god trick is part of Google's business model, bringing the relevant context to light is seen as a major threat, and thus Google fired its AI ethics team.

The firing, in addition to being an attempt to preserve the viability of the god trick, was an attempt by Google to prevent the dissolution of its accumulated power. An approach that is more comprehensive than any that has been seen so far is necessary to genuinely solve problems related to AI bias. In order to address the origins of and solutions to this problem, it is first necessary to develop a better understanding of the power dynamics within the peculiar context of the Silicon Valley tech world, and how exactly the use of the god trick came to be such an important part of the business of Big Tech.

2.3 God Tricks and Confidence Games: Objective Algorithms as Vaporware

The most successful technology companies in Silicon Valley often share a common origin story. The company was founded by a smart and ambitious person who understood both technology and business. They made many personal sacrifices to get their business off the ground. They dropped out of college and started their business in a garage. They worked their way up from nothing and single-handedly built a business empire. Stories about tech founders like Bill Gates, Steve Jobs, or Jeff Bezos all read essentially the same. This story is uniquely American, rooted in the many pre-existing American myths about the value of hard work, persistence, and ingenuity. Max Weber argued that America's particular history with both Puritanism and Calvinism helped to shape labor into an end in itself.[73] Being "hard working" thus became a core American value. This value gives labor under capitalism in America

[73]Weber, *The Protestant Ethic and the Spirit of Capitalism*, p. 26.

an "ascetic character"[74] that it inherited from its puritanical roots. That the tech founder started from nothing and suffered through it in order to become successful is thus an integral part of the founder myth, deeply rooted as it is in American culture. From the Horatio Alger myth, which teaches that anyone can be successful if they are virtuous and hard working, to the myth of the lone inventor, American history and culture is rife with examples of the "great man theory." History is often told by listing great men[75] and their achievements. The origin story of the tech founder is no different. This origin story, like the Horatio Alger tales of the 19th century, is a myth.

Oddly, one of the most common elements of the tech founder origin story is the company's humble beginning in a garage. Hewlett-Packard, Apple, Microsoft, Amazon, and Google all claim to have started in a garage.[76] One notable exception is Facebook, which didn't start in a garage, but instead started in a Harvard dorm room.[77] The garage origin story is often not at all what it seems, however. Steve Wozniak, co-founder of Apple, admitted in 2014 that, at least for Apple, "the garage is a bit of a myth; it's overblown." He goes on to explain that "we did no designs there, no breadboarding, no prototyping, no planning of products. We did no manufacturing there... there were hardly ever more than two people in the garage and mostly they were kind of sitting around doing nothing productive."[78] The story of the garage as the origin of great tech companies appears to be almost entirely fictional.

Olivia Erlanger and Luis Ortega Govela, who explore the history of the garage as a creative space in their book *Garage*, write that,

> The startup garage has become more than just a part of Silicon Valley's

[74]Weber, *The Protestant Ethic and the Spirit of Capitalism*, p. 123.

[75]I use the word "men" here intentionally. These fabricated origin stories are almost always about men.

[76]Meisenzahl, *These 5 tech companies started in garages, and now they're worth billions. These are their modest beginnings*.

[77]Ward, *Mark Zuckerberg returns to the Harvard dorm room where Facebook was born*.

[78]Steve Wozniak, *Steve Wozniak on What Really Happened in Jobs' Garage*.

> folklore; it has transformed into an image, an exportable idea that reveals a set of strategies and theorems that continue to operate in post-Fordist immaterial modes of production and consumption. Its history reveals that mythmaking has become as central to sustaining our economy as profit making. The garage became the architectural symbol that would attract the right venture capital.[79]

This mythmaking is fundamental to Silicon Valley culture, and it is perhaps the reason that tech innovators who came after Jobs and Wozniak went to such great lengths to make sure that they also appeared to start their companies in a garage. Jeff Bezos, for instance, started Amazon in a garage, but only after a brief career working for a Wall Street hedge fund called D.E. Shaw & Co,[80] and receiving a six figure investment in his fledgling company from his parents in 1995.[81] The garage was hardly necessary, but it fit with the mythology, and it could be used to attract attention from venture capital (VC). Brad Stone, author of *The Everything Store: Jeff Bezos and the Age of Amazon*, writes that "while the original Bellevue garage would come to symbolize a romantic time in Amazon's early history—the kind of modest beginnings that legendary companies like Apple and Hewlett-Packard started with—Amazon was located there for only a few months."[82]

The garage myth is really only part of the broader myth of successful innovation in Silicon Valley, one that relies heavily on the centuries-old myth of the self-made man, the lone inventor, and the great man of history.[83] It has rarely been true. It has always been part of an effort to weave together a convincing narrative. This peculiar Silicon Valley company foundation mythology is emblematic of so much of

[79]Erlanger and Govela, *Garage*, p. 80.
[80]Stone, *The Everything Store: Jeff Bezos and the Age of Amazon*, p. 19.
[81]Stone, *The Everything Store: Jeff Bezos and the Age of Amazon*, p. 33.
[82]Stone, *The Everything Store: Jeff Bezos and the Age of Amazon*, p. 34.
[83]Again, the mythology itself is gendered. The word "man" is used intentionally to highlight that these fictional stories cast men nearly exclusively as protagonists.

what makes Silicon Valley startup culture what it is. The focus, more often than not, is on the story of the company and especially the founder of the company as much as, or in some cases even more than, it is on the actual products of the company.

There is a core tension at the heart of these companies, because when the story does focus on the product rather than the founder, what the the story often promises is the development of a new technology (achievable only through the vision of the founder) that can resolve an existing social problem objectively. The history of Silicon Valley is replete with these ideas from AI and algorithmic stock trading to cryptocurrency and autonomous vehicles and weapons systems. Technology companies very frequently propose to solve a high-stakes social problem with new technology, and the technology is held up as a superior solution, despite the high stakes, precisely because it is capable of objectivity in a way that people are not.

Under the interpretation of objectivity offered by feminist philosophers of science, however, there can be no possibility of objectivity within the lone inventor / great man framework common to these companies. Donna Haraway argues that "situated knowledges are about communities, not about isolated individuals."[84] The hierarchical structure of these companies, resting as they do upon a mythology that privileges the perspective of a single individual with sole access to the supposedly true, objective vision of the future, prohibits the incorporation into science of the many limited, partial perspectives necessary for the realization of feminist objectivity. But of course these technologies are impossible — they always have been. There is no way to reach beyond the social world, with or without the aid of technology. Pretending to be able to do so is invoking the god trick. So it isn't the realization of objectivity through the invention of some fantastic new technology that is the relevant issue; instead it is the *promise* of the realization of objectivity, with that promise emerging from someone and somewhere.

[84]Haraway, "Situated knowledges: The science question in feminism and the privilege of partial perspective", p. 590.

This promise is issued through the creation of the company mythology, which is often an important part of bringing in new investors. The obvious contradiction at the heart of the mythology, that the lone inventor has brought forth an objective technology that is capable of a non-situated form of objectivity, is nevertheless situated from the standpoint of the founder himself. The founder is, somewhat ironically, embracing the view from Nowhere *from the garage*. Of course it is a contradiction, but mythologies are often full of contradictions. The point of the mythology is not logical consistency, but to identify as part of a tradition and to tell a story about that tradition. In the world of the tech startup, the narrative, above all else, is key.

There are good economic reasons for technology companies to focus their efforts so relentlessly on their image, story, and mythology, over and above everything else. Early on, small tech companies face a problem of an "information asymmetry" between the company owners and investors[85] Because the company is new and can't produce much in the way of earnings, small, innovative, new technology companies are instead valued on intangible assets such as the talent of their team or their potential for breakthroughs in research and development.[86] The predominant goal of the tech startup is thus to build faith in the future potential earnings of the company in order to attract more investment capital. A new company does not actually need to be profitable in order to be valuable, it only needs to carry the plausible potential of future profits. When these startups offer shares for sale to investors, they are not offering company earnings in the present. Instead they are offering growth potential. This potential is based on expectations of *future* earnings, and that expectation is based on narratives about the company and their potential to disrupt existing markets. In order to survive, startups must tell a convincing story about the future. Tech companies focused on growth are focused on the future, and as such have an outsize

[85]Hogan, Hutson, and Drnevich, "Drivers of External Equity Funding in Small High-Tech Ventures", p. 236.

[86]Hogan, Hutson, and Drnevich, "Drivers of External Equity Funding in Small High-Tech Ventures", p. 239.

interest in narrative. The value of the company *is* the mythology.

The prevailing attitude in Silicon Valley is one of "fake it 'til you make it." Or perhaps in the case of Facebook and others, as I will explore in Chapter 3, the attitude goes one step further to "break it 'till you make it." The implementation of the attitude of Facebook's original motto, "move fast and break things," demands a certain carelessness about one's short-term negative impact on the world. It demands that one believe fully in their plan, that despite the momentary chaos it causes, the long-run vision of the plan will eventually be realized. This pattern of behavior has been replicated by countless Silicon Valley startups, sometimes successfully, and sometimes not. The leadership of the companies that resulted in the greatest successes are touted as business geniuses, as were Steve Jobs, Mark Zuckerburg, Bill Gates and so on, while the failures are often criticized (or convicted) as frauds, as were Elizabeth Holmes[87] and Billy McFarland.[88]

The absolute, unwavering confidence that a tech CEO must have (and project) in their plan, even in the face of mountains of evidence of the foolhardy nature of their pursuit, is of the same sort as that at the heart of the confidence tricks played by successful con artists. The difference is that with a flood of VC money, tech entrepreneurs have the luxury of added time to find a way — any way — to force their vision to come to fruition. When interviewed by 60 Minutes Australia host Liam Bartlett about her failed business plans, convicted fraudster Anna Delvey[89] insisted that, had she managed to receive the 22 million dollar bank loan she fraudulently applied for, her business would have been successful and no one would have ever known that the collateral that she claimed on the loan origination documents never really existed. The exchange really is remarkable.

[87]Department of Justice, *Theranos Founder Elizabeth Holmes Found Guilty Of Investor Fraud*.

[88]Department of Justice, *William McFarland Sentenced To 6 Years In Prison In Manhattan Federal Court For Engaging In Multiple Fraudulent Schemes And Making False Statements To A Federal Law Enforcement Agent*.

[89]née Anna Sorokin

Bartlett: "I mean, you sold the story and they bought it hook, line, and sinker."

Delvey: "Maybe because the story was true. It wasn't that hard. I did have this project, and I did have this vision, and it probably would have worked out if I did have the money."

Bartlett: "But you didn't have the money! That's why the project was no good!"

Delvey: "If I would have gotten the money from the bank..."

Bartlett: "The whole Anna Delvey Foundation — your whole business — was just this house of cards. It wasn't real, Anna!"

Delvey: "So many businesses are just a house of cards. You just don't know about it."[90]

In the irrational exuberance of the tech boom of the past two decades, so much money has flooded Silicon Valley as to give even the most outrageous tech scams a decent shot at success. Anna Delvey's loan was rejected because all that would have backed it was an invented story about collateral in a non-existent German trust fund. VC-funded Silicon Valley startups' invented stories about the future, on the other hand, really are accepted as collateral. Had Delvey focused her efforts on angel investors and tech entrepreneurs in California rather than traditional banks and the art world in New York, she might have pulled it off.[91] All that is necessary for the con to be successful in the long run is that the illusion of credibility be maintained long enough for that credibility to become real — that the funding based on nothing more than outrageous storytelling holds out long enough for the company

[90]60 Minutes Australia, *How con-artist Anna Sorokin ripped off the New York elite and became a star*.

[91]Elizabeth Holmes, who did commit her fraud in Silicon Valley, built a company on false pretenses that was, at its height, valued at about ten billion dollars. (Zaw Thiha Tun, *Theranos: A Fallen Unicorn*)

to develop a legitimate revenue stream. Once this happens, it no longer matters that the company's original story didn't hold water. The many tech startups that begin this way are selling what is known as "vaporware," a portmanteau of "vapor" and "software." What they're promising to produce doesn't exist — at least it doesn't exist *yet* — and the money they bring in at the start comes from selling the story about the potential effects of the as-yet non-existent product to venture capital firms who stand to benefit in the event that the story turns out to be, or can later be made to be true. If the company is successful, the mythology becomes retroactively true.

The direction of science — the flow of information, and the course of knowledge production — is facilitated by the flow of capital. Without funding, there are no scientific projects. When the financial situation in Silicon Valley is one that is largely dictated by the whims of VC firms, the science that flows from the development of these technologies will be impacted. The biggest problem with organizing the structure of research in this way, is that attracting VC investments is not always (or perhaps even rarely) tied to the actual viability of a proposed project. What matters more than the project is the story told about the project. VC investors are buying a story about the future. There is thus an incentive among those who work in the space to focus a disproportionate amount of time and energy into the story about the project rather than the project itself. If an entrepreneur can sell the story, it doesn't matter that the product is vaporware.

This has certainly been the case in a recent high-profile example from this area: Theranos and the conviction of its CEO Elizabeth Holmes for fraud. Theranos raised VC money by selling stories about a possible future where the promised product would successfully implement the god trick, granting the owner of the technology access to perfect information via the View from Nowhere. The project was a catastrophic failure.

Theranos was a biotech startup that promised to change the world of medicine

by offering cheap, easy, diagnostic blood tests using a single drop of blood rather than multiple vials, drawn with a finger stick instead of a venous draw. The tests were to be conducted automatically in a small desktop machine called the miniLab. When Theranos' CEO Elizabeth Holmes sold this idea, the miniLab didn't exist, and prototypes suffered from major problems that engineers working on it described as insurmountable, mainly as a result of Holmes' vision of the device being no larger than a toaster, which was, according to engineers working on the project, physically impossible.[92] Pitching the project at a TED MED talk in 2014, Holmes used the word "information" twelve times in 16 minutes.[93] The main point of the pitch is that when blood tests are cheap and easy, they can be conducted frequently, providing a steady stream of actionable health information rather than annual snapshots.[94]

The fallacy behind the pitch is the implicit claim that the health information gleaned through this automatic testing is actionable without continuous oversight from medical professionals, and that the mere possession of this seemingly objective health data would be enough to influence medical interventions and change health outcomes. What Holmes was selling was the god trick. Even if the miniLab did what Holmes claimed it could do, the project would be guilty of employing the View from Nowhere, but it's much worse than that, because the miniLab never worked in the first place. Clearly what Holmes was attempting to do was the very common Silicon Valley startup trick of promising a future, world-changing moment, securing funding to make that moment happen, then attempting to find a way to make it possible before the funding runs out. This was a multi-layered fraud. The first layer is the lie about the future world-changing moment, which depends entirely on the success of the god trick. The second layer is pretending that that moment has already occurred until it does. Both layers are common to most startups, and on their own aren't

[92]Carreyrou, *Bad Blood: Secrets and Lies in a Silicon Valley Startup*, p. 97.
[93]Holmes, *TED MED 2014*.
[94]Holmes, *TED MED 2014*.

insurmountable obstacles from a business perspective.[95] The fatal flaw of Theranos, unlike successful startups like Apple, Microsoft, Amazon, and Facebook, is that what these companies promised at the outset — the product of the visionary founder that will eventually bridge the gap from mythology to reality — was improbable, but nevertheless possible. What Holmes promised was literally impossible.

I have shown that startups are focused more on building a narrative about the company than they are about building a product, and that there are good economic reasons for doing so. The idea that tech startups primarily sell a vision for a possible future rather than an existing product is an important part of my overall argument in this chapter that much of the bias problem in AI emerges from an incorrect orientation toward the function of objectivity in science more broadly. The need of tech startups to focus on the disruptive potential of innovative new technologies (thus building the case for potential future profits) has driven a related phenomenon in AI. Mathematics, statistics, and computing, lying as they do at the heart of so many consumer technologies, already structure a large part of contemporary life. In order for AI to be a technology worth investing in in the present, it has to be capable of disrupting the marketplace in the future, and thus must be capable of doing more than any existing technologies. The promise of AI, above and beyond existing technologies, is the god trick. The disruptive potential of AI, woven into the powerful narrative sold by Silicon Valley, is the possibility of finally achieving the View from Nowhere.

The false objectivity currently being sold by the various god trick AI papers described in section 2.1 is an extension of something that has been happening in Silicon Valley for a long time. These companies are not merely selling the god trick, but this is part of a pattern within the broader tech culture of selling a vision of a

[95] the obstacles are insurmountable with regard to scientific objectivity as they rely on the success of the god trick. This does not prevent them from being employed with great success in the building of a technology company.

possible future. In the same way that the garage origin story is often an attempt to create a narrative that is appealing to venture capital investors, it is the story of the god trick that matters, not its actual execution. What matters is that the potential future technology the story is about sounds just plausible enough to lure in new money. The proliferation of the View from Nowhere in AI-focused Silicon Valley startups is worse than mere mythology or even ideology; it is good for business. As Donna Haraway laments when discussing the use of artificial vision as a vector for the spread of the View from Nowhere, "all seems not just mythically about the god trick of seeing everything from nowhere, but to have put the myth into ordinary practice. And like the god trick, this eye fucks the world to make techno-monsters."[96] Silicon Valley is an incubator for techno-monsters.

2.4 AI Ethics: Reclaiming Scientific Objectivity in AI

It is impossible to get to the heart of the complex problems of AI ethics without untangling the complicated web of conflicting interests in AI science. Feenberg notes that "technology is one of the major sources of public power in modern societies."[97] An analysis of the nature and source of that power is a necessary first step before resolving the open questions regarding AI bias. One of the most visible expressions of power in the 21st century is the corporation.

The corporations that dominate the AI-science space, both in terms of the number and importance of their publications and in terms of control of large-scale computing infrastructure, have a financial interest in advancing the state of the art in AI. Simultaneously, they have a duty to return value to their shareholders. These two aims can come into clear conflict when the former impedes the latter, as apparently occurred at Google in late 2020. The power dynamic is thus tilted against scientific objectivity.

[96] Haraway, "Situated knowledges: The science question in feminism and the privilege of partial perspective", p. 581.

[97] Andrew. Feenberg, "Democratic Rationalization: Technology, Power, and Freedom", p. 706.

This situation, as computer ethicists have pointed out,[98] is in many ways analogous to the science conducted by Big Tobacco in the late 20th century. The notion that tobacco companies could conceivably participate objectively in the public health science around tobacco use is absurd on its face, and is, today, largely recognized as an irreconcilable conflict of interest. That Google would fire its entire AI ethics team rather than allow the publication of a paper that highlights the disastrous environmental effects of LLMs should be no more surprising than the conclusions reported by Big Tobacco executives, in testimony before the US congress in 1994, that tobacco is not addictive.[99]

When situations arise that position the interests of the scientific community in conflict with those of capital, corporations, as the nexus of power in the early 21st century, will always have an advantage. Google has a duty to return value to its shareholders; this duty is and always will be prioritized over and above any duties the company may have to science or academic freedom. The implications this power imbalance has on scientific objectivity are clear. The vision of a future where the View from Nowhere has been achieved technologically is more profitable than one in which it is, as it always has been, impossible. Silicon Valley corporate-sponsored AI science, not just by coincidence, but as a consequence of its fiduciary duty to shareholders, intends to sell the vision of a future AI-enabled View from Nowhere. This situation is untenable, and it is an unambiguous threat to scientific objectivity. This is not a vague or ill-defined potential future threat. The damage is real, current, and ongoing, as the firing of Google's AI ethics team demonstrates.

The fallout from the publication of the Stochastic Parrots paper[100] is the inevitable consequence of Google's endorsement of the View from Nowhere. Given its response to the allegations in the paper, Google appears not to be concerned that its LLMs

[98]Mohamed Abdalla and Moustafa Abdalla, *The Grey Hoodie Project: Big Tobacco, Big Tech, and the threat on academic integrity*.

[99]Hilts, *Tobacco Chiefs Say Cigarettes Aren't Addictive*.

[100]Bender et al., "On the dangers of stochastic parrots: Can language models be too big".

are a threat to the environment. LLMs hold great profit potential precisely because they promise to be able to see everything from nowhere, but they also result in dangerous levels of carbon emissions and energy consumption. The accumulation of these externalities is then hand-waved away as immaterial, but this dismissive attitude is only coherent if one accepts the flawed premise that it is possible to act without also being the object of action. Google's response to the publication of this paper thus demonstrates its commitment to the two core components of the god trick: to view everything from nowhere and to act without being the object of action. Timnit Gebru and the rest of the AI ethics team engaged in a good-faith effort to highlight the contradictions inherent in this view, and were fired as a result. This incident demonstrates an alarming conflict of interest: Google is one of the main players in the AI-science space, and the science, when engaged in fairly, threatens Google's business model.

When Google-affiliated authors appear on a disproportionate number of the most highly cited NLP papers, the potential for questions of AI ethics to ever see any genuine resolution is severely diminished. This is because it is necessary, in order to make progress in AI ethics, to admit that the View from Nowhere is neither possible nor desirable, but doing so will hamper not only Google's current ambitious undertaking with regard to LLMs and NLP, but the core business model of the entirety of Silicon Valley. That business model is one in which VC firms fund those companies with the most compelling stories about the future — typically those with the most convincing stories about a new technology that will achieve the View from Nowhere — with the expectation of future company growth and corresponding growth of invested capital. The problem of AI bias stems partly from power disparities established and reinforced by capitalism. These power disparities are a threat to scientific objectivity.

A potentially workable solution is to return power, as much as possible, to the workers themselves: the individual corporate-affiliated authors of those scientific pa-

pers. Scientists must not feel as if the publication of their findings will threaten their jobs. This was, in fact, part of the motivation behind granting tenure to university professors, but as more and more scientific work takes place outside of the university setting, those protections safeguard scientific objectivity to a lesser degree. These protections must be restored.

Robust worker protections, along the lines of tenure in the academic world, could be a helpful start along the road to a solution. If a research team makes a discovery and publishes those findings, a company should not be able to step in to prevent publication just because the findings are bad for business. These worker protections can and should be ensured legislatively, though a more immediate route for winning such concessions could come through an aggressive push from the workers themselves toward unionization. This puts the scientific community in the uncomfortable position of saying that the best thing for scientific objectivity is worker unionization. The discomfort of the position does not render the conclusion false however, and as I believe my argument demonstrates, it is the conclusion we must adopt.

This position is especially important for scientific research centered around AI applications that are used for high stakes decision making. The politics of the corporate-sponsored science that lie behind the development of these technologies will become, intentionally or not, integrated into the technologies themselves. The consequences of this view are many, but chief among them is that we must again abandon the View from Nowhere as a viable business, scientific, or technological objective. There are much better views of scientific objectivity offered by the feminist philosophy of science. Most of these views attempt to elevate the situated knowledge of marginalized communities, which is not possible under a strictly hierarchical power structure such as those of major corporations. Moving from the flawed view of objectivity of the View from Nowhere toward feminist objectivity requires a more horizontal power structure in order to elevate the voices of the marginalized. Greater worker protec-

tions are thus a necessary pre-condition for repairing the harm of the View from Nowhere in corporate science. Additionally, there must be an explicit acknowledgement of the social character of algorithmic decision-making from those in positions of power, including not only politicians and policy makers, but also those who design and deploy decision-making algorithms. Researchers who work on these algorithms should explicitly discuss these social effects in their published work.

The View from Nowhere is a threat to scientific objectivity. It has been adopted and tacitly endorsed by many of the largest technology companies, who are simultaneously among the most prolific and highly-cited authors of AI science papers. An analysis of the structure of funding and company growth within the broader cultural context surrounding these tech firms indicates that the acceptance of the View from Nowhere is structural, not coincidental. In order to adequately address problems of AI bias and begin to improve scientific objectivity in AI science, greater efforts at democratizing the workplace must be undertaken. Worker unionization at Big Tech firms like Google will help to improve the objectivity of AI science and counteract problems with AI bias.

Chapter 3 History, Context, and Computer Hacking

"Hacking" is a concept that is in constant flux. It changes in response to new developments and innovations in technologies, but also sometimes serves as the driver of those developments and innovations. There is a difficult-to-define relationship between science and hacking, with feedbacks between the two. The innovations of hackers often serve as an impetus to change the direction of related sciences, and that scientific output in turn shapes what is possible for hacking. Because of its unique situation within technoscience, hacking can be difficult to define, let alone to understand. In this chapter I explore the history of this concept, and in so doing, derive a workable definition, providing several examples and case studies of activities that are and are not hacking, explaining the reason for the disposition in each case. Additionally, I will explore the epistemology of hacking as a form of knowledge production and its relationship to science, particularly with regard to scientific objectivity and the capacity of non-scientist hackers to contribute objectively to science.

The novel contributions of this chapter are to (1) define the term "hacking" in light of its 60+ year history. (2) Engage in a conceptual analysis of hacking in order to draw a distinction between two common uses of the term that actually cover two distinct concepts. (3) present the view that the activities commonly engaged in by the hacker community, as illustrated by the history, are a particular form of situated knowledge production and that, as philosophers of science have argued, should be integrated with canonical forms of scientific knowledge production in order to better approach scientific objectivity.

3.1 Early Hacking: Phone Phreaking in the 60s and 70s

While on a phone call In 1957, Joe Engressia heard a faint tone in the background. He whistled along with the tone, and to his surprise the phone call immediately got disconnected.[1] Engressia, later known by the phone phreak[2] community as Joybubbles, had just accidentally discovered the secret of the 2600hz tone, the key to controlling the telephone system. Engressia suddenly became one of the world's first phone phreaks. "Phone phreaking" describes the illicit control and manipulation of the old analog telephone system. Engressia, who had perfect pitch, was blind; in 1957 he was seven years old.[3] When Engressia made his accidental discovery in 1957, very little about the inner workings of the telephone system was known by the public. However, this system was described in detail in an in-house publication of the Bell Telephone Company called the Bell System Technical Journal. Three years after Engressia pioneered phone phreaking, an article would be published in this journal that would provide future phone phreaks all the technical information they would need in order to control this system themselves. The article, Signaling Systems for Control of Telephone Switching,[4] provided details not only about the 2600hz tone, but about the new dual-tone multi-frequency (DTMF) tones used to dial a new style of phones — called "Touch-Tone" — that would be rolled out to Bell customers in 1963. The pioneering phone phreaks of the early 1960s would not only read and understand this article, but use the information contained within it to build what would later become known as the "blue box," a device used to grant to its user operator-level access to the telephone system.

[1] BBC, *A Call From Joybubbles*.

[2] The term "phone phreak" refers to a type of proto-hacker that engaged in the same set of behaviors commonly associated with computer hacking, but using the telephone and telephone network as a medium rather than computers and computer networks. This form of hacking was common from the 1950s to the 1980s. The term is a self-description chosen by the community to which it applies. The mis-spelling of "freak" is intentional.

[3] Engressia and Huse, *Joybubbles (Joe Engressia)*.

[4] Breen and Dahlbom, "Signaling Systems for Control of Telephone Switching".

These early phone phreaks, many of them college students, were not quiet about their discoveries. In November of 1963, MIT's student newspaper, *The Tech*, ran article entitled "Telephone Hackers Active" which detailed a small group's use of computers to locate available outside telephone lines. The group then, apparently for fun, tied up all available phone lines between MIT and Harvard so that the two campuses could no longer communicate by telephone.[5]

At the same time, students at Harvard were engaged in similar mischief. On May 31, 1966, *The Harvard Crimson*, Harvard University's student newspaper, ran a story describing the illegal telephone activity of four Harvard students and one MIT student.[6] The article explains that, starting in 1962, the group began experimenting with and eventually reverse engineering the telephone network in order to make free long distance calls. They kept detailed notes of their progress in a notebook labeled "Fine Arts 13" which ran to 121 pages in length.[7] Their exploits began by contacting a special telephone operator, called an "inward," an operator meant to be contacted only by other operators. As their work progressed, the five students managed to develop a device that allowed a user to place free phone calls without the need for a human operator:

> The students [discovered] that they could dispense with all operators, inward and otherwise, on calls within the United States. They merely had to beep a tone of the correct frequency into the telephone transmitter after dialing an appropriate code to connect them with a long distance trunk line. Out of $50 worth of common electronic components they constructed a device capable of reproducing tones of the 12 frequencies which are used to control automatic telephone equipment, giving them a

[5]Lichstein, *Telephone Hackers Active*.
[6]Bevard, *Five Students Psych Bell System, Place Free Long Distance Calls*.
[7]Bevard, *Five Students Psych Bell System, Place Free Long Distance Calls*.

virtual duplicate of a telephone operator's console.[8]

The device described by *The Crimson* was the blue box. This device produced the same standard DTMF tones of a touchtone phone for numbers 0-9, the pound key, and the star key, but also included a special key that produced a 2600Hz tone used by operators to route calls internally within the telephone network.[9] What these early phone phreaks had discovered through trial and error was a very serious vulnerability in the telephone system. Through experimentation and meticulous note-taking, they managed to develop their own tools to exploit this vulnerability, often for the purpose of making free long-distance calls, but also simply out of curiosity.

Perhaps the most important thing that the Harvard students discovered was the function of the 2600Hz tone, which Joe Engressia had discovered by accident nine years before the publication of the *Crimson* article. The Harvard students observed that the introduction of the 2600Hz tone disconnected their calls, but after more trial and error, realized that while they had been disconnected from the other party to whom they had been speaking, they weren't disconnected from the telephone network entirely. The strange thing about 2600Hz is that it disconnects one side of a phone call, but not the other. The purpose of this behavior was to signal to the other side of the call that the line was now idle, i.e., that the other party had hung up their receiver. This was an important part of the way that the telephone system worked in the USA up until the late 1980s. During this time, when a phone call was placed, both sides of the call would be physically connected by a switch (or more likely, a long series of switches) forming a circuit. The circuit, starting with the phone that placed the call, was routed through a local telephone company central office, called an "exchange"[10] where it would activate a switch that completed the circuit

[8]Bevard, *Five Students Psych Bell System, Place Free Long Distance Calls*.

[9]E. Goldstein, *The Best of 2600: A Hacker Odyssey*, p. 24.

[10]the exchange was common for all telephone numbers in the same region. The exchange is more localized than the area code. For instance, the 795 exchange in Sykesville, Maryland is in the 410 area code for Baltimore, Maryland. So a phone number such as "410-795-1254" can be broken down

by connecting the call to the dialed number. When a 2600Hz tone was introduced to the line during a call, the call would be cut off — that is, the other party would be removed from the circuit — but the circuit from the originating phone into the local exchange remained active.[11] At this point, with the originating telephone still connected to the central office, another number could be dialed, and a new call placed without ever hanging up. If a call were placed and connected, and a 2600Hz tone introduced and then removed, the receiving end of the call would hear the tone, recognize the idle state of the line, and disconnect, at which point the caller would still be connected to the central office, but not to a receiving line. The exchange to which the caller was connected would simply wait for a new number to be dialed. The telephone system could be abused by, for instance, placing a toll-free call to a far away place so that the call would be routed through a long-distance trunk line and connected to a distant exchange. If, after being connected, a 2600Hz tone were introduced, the receiving line would be disconnected, but the sending line would still be connected to the distant exchange via a long-distance trunk line. Because the local exchange recorded that the outgoing call was placed to a toll-free number, it wouldn't be billed to the customer who owns the sending line, but after disconnecting the call with the 2600Hz tone, the call originator would be free to dial a new number as if it originated from the distant central office to which they were still connected. If this new number were local to that exchange, that call wouldn't incur any billing either, so the 2600Hz tone could effectively be used to make free long-distance phone calls. Historian of phone phreaking Phil Lapsley summarizes the core vulnerability in the mid-20th century American telephone system very succinctly with three main points,

> "The first was that sending a 2,600Hz tone down the telephone line resets

into three parts: the area code (410), the exchange (795), and the line number (1254).

[11] Breen and Dahlbom, "Signaling Systems for Control of Telephone Switching", p. 1422.

> the remote switch but doesn't affect the local switch. The second was that you could then reroute a phone call from the remote switch to whereever you want. And the third was that the local switch is in charge of billing, so it continues to bill you for whatever call it thinks you originally made."[12]

While it took the talented engineers at Bell Telephone, Bell Labs, and Western Electric the better part of a century to build this system, it only took the students at Harvard and MIT a few semesters to reverse engineer it and start using it for their own purposes. In a presentation on the history of phone phreaking given at the HOPE hacker conference in 2012, historian of phone phreaking Phil Lapsley showed a slide of an old black rotary phone that had been seized by the FBI as evidence in a law enforcement crackdown. Pointing to the phone, he described the historical significance of a plain telephone receiver as a piece of evidence in an FBI investigation, and in doing so, captured something essential about the spirit of phone phreaking that became a core element of the hacker mindset,

> "to me this represents not a black rotary phone but it represents curiosity personified. Because if you have a certain mental defect like many of us do, you look at something like that and you say, 'wow! I wonder how that works. I wonder where it goes to. I wonder what happens if you take it apart. I wonder what happens if you use it in ways that it was not intended to be used. I wonder what happens if you dial numbers that are not in the phone book. I wonder what happens if you play with it. Where do those wires go? How does that all work?'"[13]

This spirit of wonder and curiosity is a defining characteristic of the pioneering phone phreaks of the 1960s and 1970s. This, combined with a carefree attitude about

[12]Lapsley, *Exploding the Phone: The Untold Story of the Teenagers and Outlaws who Hacked Ma Bell*, pp. 55–56.

[13]Lapsley, *Phone Phreak Confidential: The Backstory On The History Of Phreaking*.

consequences and perhaps some disdain for the law, was what helped make phreaking (and later hacking) into a social force rather than just a one-time historical curiosity that would lose its significance when the phone company fixed these vulnerabilities. Lapsley continued,

> "I've talked to so many of these phone phreaks and they would tell me, 'we're 15-year-old, 16-year-old, 17-year-old kids we don't have anybody to call — right? — we don't know anybody. The only thing we want to do is, we want to call — you know — we want to call non-working test numbers, we want to see how close we can get to the North Pole, we want to — you know — we just want to see other networks. We're just curious!'"[14]

Important to note in these descriptions of phone phreaking is that the problems that phone phreaks were engaging with were problems that presented themselves in daily lived experience. The desire to explore the telephone system was a result of interacting with it, recognizing some interesting characteristics, and making a conscious decision to pull on that thread and see what happens. Hacking is a mode both of engagement with one's environment and relationship with other users of a technical system. That the problems that confronted the phone phreaks emerged out of life is important because it stands in contrast to what would later characterize the hacking mentality in the 2020s — a mode of engagement that is significantly less connected to a web of contextual relations. When the phone phreaks were exploring the phone system, they were, in part, exploring their own social world.

The discovery of the function of the 2600Hz tone allowed phone phreaks to do more than just make long distance phone calls. It allowed for the creation of one of the first virtual spaces where phreaks from across the United States and across the

[14]Lapsley, *Phone Phreak Confidential: The Backstory On The History Of Phreaking*.

world could meet and talk not only about technology and their relationship to it, but about themselves and each other. This was made possible through the accidental discovery of so-called "party lines."[15] In the early days of the telephone system, before it was possible to connect every subscriber to their own dedicated phone line, it was necessary for a large number of households to share a single phone line, called a "party line." When a party line subscriber picked up their receiver, they may discover that their neighbor is already on the line.[16] The "party lines" discovered by 1960s and 1970s phone phreaks were named identically, but do not describe the same system. The party lines used by phone phreaks as the first public virtual meeting spaces were special, often inoperable, telephone numbers. Upon dialing such a number, a caller would often hear either a busy signal or a recorded message from the telephone company saying something like, "the telephone number you have dialed is out of service. Please hang up and consult a directory." Phone phreaks discovered that if two people called such a number at the same time, in some rare cases they could talk to each other either over the sound of the busy signal, or in the silence between the end of one iteration of the voice recording and the start of the next.[17] This trick didn't work for every inoperable phone number, but phreaks quickly found which ones did work, and these numbers became hang-out spots for phreaks around the country. If a particular party line number was long distance, a phreak could still access it for free by using the 2600Hz tone to break out of a toll-free call to the exchange local to the party line before connecting to it. Included among the phone phreaks that have stated that they got their start in the world of phreaking with the use of party lines

[15] also sometimes called "conference calls"

[16] and could, in fact, eavesdrop on their conversation if they chose to.

[17] Doorbell, *How I Became a Phone Phreak: The Dark Side of 'Party Lines,' November 1970*.

are Denny Teresi[18] and Jim Fettgather,[19] both friends of Joe Engressia, and both also blind. All three made frequent use of party lines in the Los Angeles and San Jose area in the 1960s.[20]

It seems unlikely to be a coincidence that Engressia, Teresi, and Fettgather (along with many other extremely successful phone phreaks) were blind.[21] The telephone presented a unique social opportunity to blind people that was right in front of sighted people all along, but did not stand out in the way that it did to Engressia and others due to its obviousness. It was the fact that the telephone represented, for Engressia and others, a mode of interaction with the social world that so changed their relationship with it that they spent so much time on the phone. The problems they were able to identify and engage with were social problems. They were engaged in a philosophy of sorts simply by directly confronting these problems of life in the way that they were able to in their time. The fact that so many early phone phreaks were blind is philosophically important because it highlights the fact that these phreaks were directly engaged with the problems of life. One such problem was and remains the pervasive inaccessibility of so much of our infrastructure to blind people. In

[18]In a 2014 interview(Draper, *Crunch Life 01: Denny Teresi*), John "Captain Crunch" Draper recounts the story of how he met Denny Teresi. According to Draper, he made a call using ham radio asking for anyone listening to call him on his telephone. He received a phone call from Teresi, who provided a callback number for a party line (as Teresi didn't want to reveal his real phone number). When Draper called in to the party line and made contact with Teresi, they eventually became friends, and Teresi showed Draper how to use the 2600Hz tone, which Teresi had been generating with an electronic organ synthesizer connected to the phone line. This inspired Draper to build a blue box to generate the MF tones necessary to accomplish the same task in a much smaller footprint.

[19]Lapsley, *Exploding the Phone: The Untold Story of the Teenagers and Outlaws who Hacked Ma Bell*, p. 147.

[20]Lapsley, *Exploding the Phone: The Untold Story of the Teenagers and Outlaws who Hacked Ma Bell*, p. 147.

[21]Joe Engressia is certainly the most famous blind phone phreak, but there have been a surprising number of blind phone phreaking pioneers, including Bill Acker(Lapsley, *Bill Acker, 1953-2015*) and other members of the '2111 gang' of phone phreaks, Denny Teresi and Jim Fettgather(Lapsley, *Exploding the Phone: The Untold Story of the Teenagers and Outlaws who Hacked Ma Bell*, p. 164), Rick Plath(Lapsley, *Exploding the Phone: The Untold Story of the Teenagers and Outlaws who Hacked Ma Bell*, p. 149) Muzher, Shadde, and Ramy Badir(Kaplan, *Three Blind Phreaks*), and Matthew Weigman(Poulsen, *FBI Charges Blind Phone Phreak With Intimidating a Verizon Security Official*)

a world built for sighted people, using technologies built for sighted people, they reimagined, remade, and transformed these technologies in direct response to these problems.

Dilthey's work is relevant to understanding the context both of problems as they are encountered in life and solutions as they are devised by science. Two of Dilthey's ideas are directly relevant to the philosophy associated with phreaking and hacking, the "standpoint of life" and the "observational standpoint." Robert Scharff describes the standpoint of life as "the standpoint of directly experienced 'historical human life' as we live through it — and thus also as the standpoint adopted by human scientists, who want to understand this life 'in terms of itself.'"[22] Dilthey is adamant that science is a social practice, that "all science is interpretive and all interpretation is contextual."[23] When taken from the standpoint of life, the form of technologically enabled social problem-solving engaged in by late 1960s and early 1970s era phone phreaks (particularly those who were blind) is a human problem. Social scientists interested in solving the problem that spurred Engressia and others to action could themselves have attempted to devise solutions, but to do so they would have had to begin by separating themselves and their work from the world by taking the "observational standpoint." When approached through the observational standpoint, the blind phreaks seeking social connection through the telephone cease to be human beings living through a particular form of social isolation, a situation from which they are seeking relief, but instead become variables in an equation abstracted away from their lived experience. The importance of preserving context and taking the standpoint of life in this example should be clear — by failing to foreground social, cultural, and historical context, we risk losing sight of the goal which initially gave us the reason to pursue an abstract mathematical-technological solution in the first

[22]Scharff, *How History Matters to Philosophy: Reconsidering Philosophy's Past After Positivism*, p. 161.

[23]Scharff, *How History Matters to Philosophy: Reconsidering Philosophy's Past After Positivism*, p. 157.

place. The choice is between using the abstract to make changes to the real or allowing the real to become subsumed by the abstract. Taking the standpoint of life protects us from the latter.

It was with the same goal of developing closer social bonds that Facebook opened its social network in 2004. Facebook's original mission statement was "making the world more open and connected."[24] The major difference between the approach toward solving this problem taken up by the early hackers and that of Facebook is that Facebook was forced to abstract away from the life problems that initially gave rise to their intervention. Their relationship to the problem of social isolation was clouded by the need to bring new users to an ad-supported social network. This further separated the researchers, programmers, and business people engaged in the problem solving activity from the problems of life. As a result, even if the solutions proposed by a corporate entity like Facebook do manage to make the world "more open and connected," they won't stand as genuine solutions, because those solutions are not appropriately grounded in life and in lived experience. This appears to be a deficiency that even Facebook itself was aware of, as TechCrunch pointed out in 2017 when Facebook abandoned the original mission statement,[25] As TechCrunch noted after the change, "'making the world more open and connected' had one fundamental flaw: it didn't push for any specific positive outcome apart from more connection."[26]

Scharff describes the standpoint of life as an attempt to "understand life on its own terms."[27] In contrast to Facebook and other large institutional players who mimic the hacker methodology without replicating its orientation toward life, the solutions arrived at by the early hackers represent a genuine attempt to understand life 'on its

[24]Hoffmann, Proferes, and Zimmer, ""Making the world more open and connected": Mark Zuckerberg and the discursive construction of Facebook and its users".

[25]as of this writing, the current mission statement for Facebook, which has since changed its name to Meta, is to "give people the power to build community and bring the world closer together."(Meta, *Our Mission*)

[26]Constine, *Facebook changes mission statement to 'bring the world closer together'*.

[27]Scharff, *Heidegger Becoming Phenomenological: Interpreting Husserl through Dilthey, 1916–1925*, p. 13.

own terms' because both the problems and solutions that they engage with emerge from life. This orientation toward life and the problems that emerge out of it is at the core of what makes hacking what it is, and why, when the same methodologies are taken up by large institutional forces later, they fail to achieve similar results.

Through the use of this methodology, what the hackers demonstrate — convincingly — is that technology truly can be, as Feenberg believes, transformed.[28] When approached from the standpoint of life, if technology isn't working the way one believes it should, it can be remade and reworked into something new and different. Hacking is a direct expression of David Graeber's statement that "[the world] is something that we make, and could just as easily make differently."[29] It is a form of direct action that seeks to repurpose existing technology toward solving the problems of life. Reclaiming agency over technology and actively asserting the role of the human as the author of the output of technology is exactly what Joe Engressia was doing as a teenager when he was exploring the Bell telephone system. There were no virtual social spaces,[30] but Engressia, Teresi, Fettgather, Acker and others decided to create one. They didn't have to invent new technology in order to do it. They only had to creatively re-invent the world of possibilities for an existing technology. The telephone was designed for a specific use, but in reinventing the possibilities for its use, Engressia was in effect saying that the phone itself can't tell us how to use it.

3.2 The History of the Term "Hacker"

The contemporary understanding of the term "hacking" or "hacker" has its origins in the mid-20th century confluence of multiple discrete social, political, and technological movements, as well as many historical contingencies, chance discoveries, and odd coincidences. The earliest known printed usage of the term "hacking" with its

[28] Andrew Feenberg, *Transforming Technology: A Critical Theory Revisited*.

[29] Graeber, *The Utopia of Rules: On Technology, Stupidity, and the Secret Joys of Bureaucracy*, p. 89.

[30] and this was especially so for the blind, who had a greater need for them than sighted people

contemporary meaning was in the November 20, 1963 edition of *The Tech* which contained an article titled, "Telephone Hackers Active."[31] Hacking and phone phreaking have always been part of the same interwoven history; phreaking really is just a subspecies of hacking. Both are engaged in the same activity, but phreaking refers to hacking limited to telephones and the telephone network.

The concept of hacking and the use and meaning of the term has as long and winding a history as hacking itself. Many outside of the hacking world may, even today, have a definition of hacking that is wrapped up in notions of criminality. Recent descriptions of computer viruses, ransomware, and various other technologically enabled extortion scams are often described as being perpetrated by a "hacker." Legislators in the United States Congress have taken these types of criminal threats seriously and have held hearings on the issue of criminal hacking. For example, on November 16, 2021, the House Committee on Oversight and Reform held a hearing called, "Cracking Down on Ransomware: Strategies for Disrupting Criminal Hackers and Building Resilience Against Cyber Threats."[32] While this usage of the word "hacker," was carefully qualified by the House as "criminal," its connection to criminality and "cyber threats" is unmistakable. The co-occurrence of these terms is common, and appears to have given rise to a common understanding of the word "hacking" as being related to computer crime. Hackers, in the popular imagination, are criminals who illegally gain access to computer systems in order to steal, blackmail, scam, or otherwise cause harm to their innocent victims. This is not how I use the term here. The history of hacking that I engage with in this chapter may, at times, involve criminal activity, but this is not the defining characteristic of the historical movement I intend to capture. The hackers under study here generally do not have malicious intent, and are instead simply interested in learning how techno-

[31]Lichstein, *Telephone Hackers Active*.

[32]Maloney, *Cracking Down on Ransomware: Strategies for Disrupting Criminal Hackers and Building Resilience Against Cyber Threats.*

logical systems work through experimentation. Richard M. Stallman, known by the hacker and free software community as "RMS,"[33] an early pioneer of free software, describes hacking as a set of activities that have in common "playfulness, cleverness, and exploration." According to RMS, "hacking means exploring the limits of what is possible, in a spirit of playful cleverness."[34] In the hacker community, this definition is generally accepted, and it is the basis of the definition of the term as I will use it here.

There is a distinct, yet related understanding of hacking that has much more political undertones, and is related to political agitation in the US surrounding the Vietnam war. This can be traced back to Abbie Hoffman's *Steal This Book.*[35] One of the first histories of the hacker movement was authored by Bruce Sterling in 1992, and was released initially as a traditional book, but later as one of the very first ebooks. Sterling mentions the connection between Hoffman and phone phreaking relatively early in the book,

> Hoffman, like many a later conspirator, made extensive use of pay-phones for his agitation work — in his case, generally through the use of cheap brass washers as coin-slugs. During the Vietnam War, there was a federal surtax imposed on telephone service; Hoffman and his cohorts could, and did, argue that in systematically stealing phone service they were engaging in civil disobedience: virtuously denying tax funds to an illegal and immoral war.[36]

Hoffman's interaction with the telephone system was not, as Stallman argued was core to hacking, an effort to engage in playful cleverness. His use and abuse of pay-

[33]I am including Stallman's contributions to hacking here, but it should be noted that there have been many credible accusations of inappropriate behavior against him, and in 2019 he was forced to resign from both MIT and the Free Software Foundation after publicly making offensive comments about rape(Lee, *Richard Stallman leaves MIT after controversial remarks on rape*)

[34]Stallman, *On Hacking*.

[35]A. Hoffman and Fithian, *Steal This Book (50th Anniversary Edition)*.

[36]Sterling, *The Hacker Crackdown*.

phones was political. It was an act of anti-war activism meant to steal long distance phone calls with the dual purpose of furthering the aim of anti-war activist organization, and depriving the federal government of tax revenue to fund the war. Hoffman wrote in *Steal This Book* that "ripping off the phone company is an act of revolutionary love."[37] This antagonistic orientation toward both the state, and the coercive power structures embedded in, of all places, the long distance telephone network, portends another element which I believe is core to the hacker mentality. That is that hackers act as if, in most of the things they do, they are standing in opposition to coercive hierarchies and working towards a more free world. There is an undeniably political element to the hacker subculture. This element prizes individual freedom and autonomy, and opposes any form of coercive hierarchy. It has a clear libertarian streak, but was in the early days agnostic with regard to left or right leaning libertarianism. This political element of the hacking culture, to this day, remains embedded in the now unambiguously right-leaning libertarianism that is extremely common in Silicon Valley.[38] Hoffman believed his actions were part of a broader civil disobedience campaign; he saw himself as standing in opposition to government and the machinery of war in the same way that later hackers would see themselves as standing in opposition to similarly unjust entrenched power structures[39]. Importantly, seeing the power dynamic in this way provides a kind of built-in justification for maintaining a casual indifference toward the law. If law-breaking is civil disobedience, it is justified, so ignoring the law is not a problem. One consequence of Hoffman's involvement in this space is the later framing of hacking as a form of civil disobedience. This deep-seated relationship between hacking and civil disobedience is likely one major reason for the continued associations between hacking and criminality.

[37] A. Hoffman and Fithian, *Steal This Book (50th Anniversary Edition)*, p. 75.

[38] see for instance Barbrook and Cameron ("The californian ideology") and Golumbia (*The Politics of Bitcoin: Software as Right-Wing Extremism*)

[39] for instance, the portmanteau "hacktivism" comprised of the words "hacking" and "activism" came into common use in the late 2000s and early 2010s due to the actions of individuals and groups including Chelsea Manning, Wikileaks, Edward Snowden, Anonymous, and LulzSec.

Hoffman's influence was not limited to the political character of the hacking movement, but also served to shape the style of many of the primary source writings and contemporaneous written records created by many hackers through the 1980s and 1990s. Jason Scott, an archivist and historian of 80s and 90s Bulletin Board Systems (BBSes) who maintains the well-known textfile archive textfiles.com, claimed in a 1999 speech at DefCon 7 that,

> To me, the original textfile is this, Abbie Hoffman's "Steal This Book". When I leaf through this book, which was printed in 1970, I get a very very strong sense that I get reading a classic textfile. Abbie's book is chapter after chapter on how to get free medical help, free food, and how to rip off the phone company, supermarkets. Basically, it's a how-to manual on fighting and using the system on your way to revolution. Abbie presents his ideas in a witty style, with lots of examples, lots of encouragement, and from a position where he encourages the whole thing as a big game and the Right Thing to do. And when you look over a lot of textfiles that were written, a lot of them take that tack: here's some information I've gotten, it's given me lots of fun, I've benefited from it, and I want you to benefit too. How-to's written by computer telling you how to get by in the world, occasionally at someone else's expense.[40]

Hoffman was also responsible for co-founding, along with Albert Bell, what was likely the first phreaking magazine, *The Youth International Party Line* (YIPL) in 1971. The very first line of the first issue, dated June 1971, reads, "We at YIPL would like to offer thanks to all you phreeks out there."[41] This is an early written instance of the nickname given to hackers of the telephone system, later given the standardized spelling "phreaks."

[40] Scott, *The Text of my 1999 DEFCON 7 Speech*.
[41] Abbie Hoffman, *Youth International Partyline Newsletter*.

A final consideration for a component of the full meaning of "hacking" should include some form of generalized good-natured mischief and light chaos. This should be qualified, however — any mischief or chaos, in order to be considered hacking and not mere trouble making, should require a high degree of technical proficiency and should serve as a proof of concept to other hackers that the perpetrator of the mischief or chaos in question was able to pull it off due to their skill, cunning, and determination. In this instance, the chaos is not the primary goal. The goal is to get others to notice the technical achievement, and the chaos is a mere means to that end. It is not malicious or mean spirited. It is not meant to destroy property or cause financial damage; it is meant to cause others to take notice. This is, in effect, a form of high-tech bragging. After pulling off a really clever or particularly difficult hack, hackers love to brag. This isn't motivated by ego per se, but it is an extension of the "curiosity" element of hacking — hackers love to share what they've done for the same reason academics write papers. It is helpful and productive to discuss new and interesting ideas with one's colleagues. In the early days at least, there were no journals where a hacker could publish their findings[42]. But if those findings were to be shared, it would often require demonstrating the newfound knowledge on production systems. A possible example[43] of this type of behavior is the infamous Max Headroom Broadcast Intrusion. On November 22, 1987, someone broke in to the broadcasts of two local Chicago television stations, WGN,[44] and WTTW,[45] with a bizarre video featuring a person wearing a Max Headroom mask. Al Skierkiewicz,

[42]Though there certainly are venues for this today. The premiere conference for hackers is DEF CON, held annually in Las Vegas, Nevada, and typically drawing over 30,000 attendees from information security, signals intelligence, and other computer security oriented professional fields. DEF CON has been held every year since 1993. Other, similar conferences include HOPE, Shmoocon, Blackhat, and the Chaos Computer Club.

[43]I say possible example because in this instance the identity of the hacker, and thus the nature of their motivations, remains unknown.

[44]The Museum of Classic Chicago Television, *WGN Channel 9 - The first Max Headroom Incident*.

[45]The Museum of Classic Chicago Television, *WTTW Chicago - The Max Headroom Pirating Incident*.

a WTTW broadcast engineer at the time of the signal intrusion, believes that the unknown perpetrator must have been "a broadcast engineer, a satellite engineer, or a ham radio operator, and probably a combination of at least two of those in order to pull this off."[46] While the true motives of the person who conducted the broadcast intrusion are unknown, one plausible explanation for why they did this is the simple fact that they could.

Providing a standardized definition of the term "hacker" is difficult because hacking is a moving target. Given that at least one core component of the meaning of the term relates to working with and relating to technology, and during the time in question, the technology in use was in a state of rapid change, this alone would seem to stymie any effort of nailing down a single definition that is consistent across time. Suffice it to say, any good definition of hacking must combine in some indeterminate ratio, the elements of Richard Stallman's notion of "playfulness, cleverness, and exploration" with the political element contributed by Abbie Hoffman of resisting coercive or unjust hierarchies and generally ignoring the law in doing so, as well as some measure of plain braggadocio and general mischief making.

Hacking is a form of applied problem solving, generally centered around technology. Additionally, hacking is characterized by (a) "playfulness, cleverness, and exploration"[a] (b) resistance to coercive or unjust hierarchies and a general disregard for the law should it conflict with a discovered solution (c) braggadocio and general mischief making

Figure 3.1: A tentative definition of hacking

[a]Stallman, *On Hacking*.

Hacking is a concept that has been in flux since the 1960s. It has changed with the times and with the technology it makes use of; it has changed in response to the social and political climate; it has changed in response to new forms of mass media and increasing rates of participation in those media, but it all started with phone

[46]Shefsky, *30 Years Later, Notorious 'Max Headroom Incident' Remains a Mystery*.

phreaking. The basic components of the definition above remain in place to this day, but as advances in consumer technology changed the hacking landscape from the late 1970s through the 1990s, the definition of hacking would change as well.

3.3 Computer Hacking and Hacker Culture in the 80s and 90s

Hacking changed significantly in the late 1970s and early 1980s as computers became available to the general population. The first such computers were sold as hobbyist kits like the Altair 8800, released in 1974,[47] but later pre-assembled systems became available including the Apple II (1977), Tandy TRS-80 (1977), Commodore PET (1979), Atari Model 400 and 800 (1979), IBM PC (1981), and Commodore 64 (1982).[48]

These newly available computers were very quickly used in conjunction with existing phone lines in order for computer users to communicate electronically. In 1978, Randy Seuss and Ward Christensen launched the very first electronic bulletin board system (BBS) accessible by the public.[49] This was a dial-up service that home computer users could access with a modem. A typical BBS contained message boards that users could post to and sometimes a files section where users could share files with one another. As home computers and modems became more accessible and popular, BBSes[50] grew in popularity as well. Before the Internet was available to the public, the BBS was the primary means of electronic communication via computers.

Some of the most important publications related to hackers and hacking at this time centered much of their writing around both phone phreaking and the BBS. The BBS became a meeting place for like-minded people, much like the so called "party lines" used by phreaks in the 1960s and 1970s. The first article in the first

[47]Smithsonian National Museum of American History, *Altair 8800 Microcomputer*.
[48]Museum, *Timeline of Computer History*.
[49]Gilbertson, *Feb. 16, 1978: Bulletin Board Goes Electronic*.
[50]The acronym "BBS" is difficult to pluralize. I am using "BBSes", which is the most common pluralization

issue of *2600 Magazine* (January 1984) begins with a discussion of a BBS called OSUNY,[51] popular with hackers and phone phreaks in the early 1980s. Similarly, the first issue of *Phrack Magazine* (November 1985) discusses phone phreaking methods and encourages readers to share this information on their local BBS.[52] Years later, Bruce Sterling would spend significant time discussing the importance of the BBS to the hacker community in his 1993 history of hacking, *The Hacker Crackdown*. Sterling writes of the BBS,

> Boards cover most every topic imaginable and some that are hard to imagine. They cover a vast spectrum of social activity. However, all board users do have something in common: their possession of computers and phones. Naturally, computers and phones are primary topics of conversation on almost every board. And hackers and phone phreaks, those utter devotees of computers and phones, live by boards. They swarm by boards. They are bred by boards. By the late 1980s, phone-phreak groups and hacker groups, united by boards, had proliferated fantastically.[53]

3.3.1 Textfiles as Hacking

The rise in the popularity of the BBS led to the proliferation of "textfiles,"[54] sometimes spelled "filez" or "philez" and often referred to as G-files for "general files" as this is the section of most BBSes where they were posted. *Phrack Magazine* was released in this format and was shared widely. These text files were responsible for the cultivation and development of much of what is understood as hacker culture

[51] Emmanuel Goldstein, "Ahoy!"

[52] King, "Introduction..."

[53] Sterling, *The Hacker Crackdown*.

[54] written as the compound word "textfiles" (with no space) intentionally as a way to distinguish the types of files that spread on 1980s BBSes from an ordinary plain-text file. "Text file" refers to any file saved in plain-text (.txt) format on a computer. "textfile" refers to a type of file with particular cultural significance to the history of hacking, commonly found on BBSes in the 1980s and beyond.

today. Many of these textfiles now serve as the best primary source historical material documenting this period in the history of hacking. Textfiles covered a vast array of topics, but generally centered on the cultural and technical elements of hacking. Textfiles document what it was like to be a hacker in the 1980s. In this era, there is ample evidence that hackers themselves struggled to capture the essence of what it meant to be a hacker. Many textfiles are merely an attempt to define hacking, identify and discuss the shared experiences common to the hacking community, and catalog the types of activities that hackers were engaged in. *2600 Magazine*, to this day, regularly runs a column called "The Hacker Perspective" which is simply a collection of articles written by various authors describing their history with hacking and explaining what hacking means to them. In aggregate the column can be seen as a longstanding attempt to define hacking. When the editors of *2600* called for new write-in submissions for this column in 2018, they asked that submissions answer the questions, "What is a hacker?"; "How did you become one?"; "What experiences and adventures did you live through?"; "What message can you give to other aspiring hackers?"[55]

A particularly salient example of a textfile that attempts to define hacking and answer questions about the common experiences lived through by hackers is "The Hacker Manifesto," published originally in *Phrack Magazine* in 1986 under the title "The Conscience of a Hacker"[56] by the pseudonymous author, The Mentor.[57] The manifesto, which was shared widely on BBSes, describes the frustrations of being a motivated, technologically competent autodidact in the American public school system. This document is less of a manifesto and more of a written remonstration expressing young hackers' collective frustration at having their creativity actively stifled by teachers and administrators at school. Because many hackers at this time —

[55]2600 Magazine, *Hacker Perspective Submissions*.
[56]Blankenship, *The Conscience of a Hacker*.
[57]Born Loyd Blankenship

and today for that matter — were (are) young people discovering for the first time the new possibilities awaiting them through their creative reinvention of a technology's purpose, a very common theme among primary source accounts of the phenomenon of hacking is a long list of complaints about the oppressive nature of the traditional institutions that rejected this methodology as illegitimate. A common experience among hackers was a strong sense of rejection by mainstream institutions of their creative and effective, if unorthodox, methodology. Often when a young hacker shared an accomplishment they were proud of with a trusted adult at school, they were met with immediate rejection, castigation, and even punishment. The Hacker Manifesto concludes, "yes, I am a criminal. My crime is that of curiosity... My crime is that of outsmarting you, something that you will never forgive me for."[58] The same sentiment is expressed by many other classic textfiles from the same time. A textfile called "The SchoolStoppers' Textbook" begins by imploring readers to, "Liberate your life - smash your school! The public schools are slowly killing every kid in them stifling their creativity and individuality making them into non-persons. If you are a victim of this, one of the things you can do is fight back."[59] A textfile called "Screwing with School Computers" begins by explaining that "Hacking is all about information. To become a hacker you must learn everything you know on your own or by listening to other hackers. Schools and what they call 'education' has little to do with learning. So this file is here to show you some truly productive things to do at school."[60]

What many of the 1980s textfiles reveal about hacking is that it exists primarily in distinction from mainstream, institutional methods of education and problem solving. It depends upon the creativity and agility of individual problem solving. Because hacking depends on the creative, playful reinvention of existing technology for the purpose of solving lived problems, the institutional uptake of those solutions

[58]Blankenship, *The Conscience of a Hacker*.
[59]Youth International Party, *SchoolStoppers Textbook*.
[60]Liquid Bug, *Screwing with School Computers*.

immediately removes them from consideration as hacking. Once institutional uptake has occurred, those solutions no longer stand in distinction to mainstream solutions, because they *are* the mainstream solutions. They are no longer creative reinterpretations, because they have become orthodox interpretations. Successful hacking is allopoietic. That is, when a creative hack is particularly good, it will be recognized by the non-hacking world as a canonical solution. Once a solution derived through hacking becomes standard, it is no longer hacking. This may appear at first to be a contradiction; successful hacking essentially subverts itself. The better the hacker's solution, the more likely it is to become institutional, and thus, at odds with the goals of hacking. The view of hacking as allopoietic, however, defuses this apparent contradiction. Hacking is, at its root, a creative endeavor. It is a means by which one creatively reimagines the use of something in order to solve a problem. The solution arrived at through this creative reimagining can then be packaged, distributed, and imported into new contexts.

Once a problem has been solved, there is no need to reimpliment a custom solution to the exact same problem each time it comes up. There's no need to reinvent the wheel, as the saying goes. Solutions are portable. A solution derived by hacking is not trapped within the hacking ecosystem. Non-hackers are free to use the same solutions. The core elements of hacking: creativity, cleverness, and playfulness, can exist within the first derivation of a unique solution to a problem without existing in all subsequent applications of the same solution. The hacking occurs in the derivation of the solution, not in its later application. The textfiles that emerged out of this era that proposed unique creative solutions to common problems were themselves a form of hacking even if the application of the information they contain is not.

What the 1980s textfiles demonstrate is that hackers often define themselves in distinction to non-hackers. So hacking is adversarial in the sense that to be a hacker is to exist and operate outside of the bounds of traditionally accepted problem-solving

spaces. Additionally, these textfiles show that even though this adversariality exists in the generation of solutions, it need not exist within the solutions themselves. Though traditional and institutional systems reject the problem-solving methods employed by the hackers, they do not always reject the solutions derived by this methodology. The hacking resides in the problem-solving, not in the solution.

3.3.2 BBSes as Hacking

Not everyone who used a BBS was a hacker or a phone phreak, though there was a significant overlap in these communities. Jason Scott of textfiles.com has archived 771 files related to phone phreaking saved from BBSes in the 1980s or early 1990s[61] At the same time, the breakup of Bell Telephone in 1984 along with technological advances in telephone switching[62] signaled the beginning of the end of the phone phreaking era.

Even as phone phreaking became less common, with the growth in popularity of computing and an explosion in other new consumer technologies, there was no shortage of hacking conversation topics on BBSes. In the 1960s and early 1970s there were relatively few complex telecommunication or computational systems for hackers to explore; they were limited more or less just to the telephone network and perhaps ham radio. With the availability of computers, modems, BBSes, and later the World Wide Web, the surface area available for exploration by those who embodied the hacker mindset grew exponentially.

A local dial-up BBS allowed communication with anyone else able to dial in, which meant that most BBS users were in the same area code as the BBS. The alternative was to pay for a long distance phone call in order to dial in, which was

[61] Scott, *Phone Phreaking*.

[62] In the 1980s the telephone system transitioned from in-band to out-of-band signaling. This separated a single channel for both communications data and system control data into two distinct channels. This transition meant that the 2600Hz tone no longer allowed a phreak to seize a trunk line

often prohibitively expensive. This was especially the case after blue boxes stopped working and getting access to free long distance calls was no longer as easy as it had been in the past. Between the breakup of Bell in 1984 and the widespread availability of the public Internet in 1993, nationwide electronic communication via BBS was possible, but sometimes difficult.

One way that BBS users were able to communicate with one another nationally was by networking the BBSes rather than the users. An example of such a network was Fidonet, which began in 1984 and networked over 20,000 individual BBSes worldwide, allowing users to send and receive email with other users on the network.[63] There were many other systems like this, including Usenet and Darpanet which were available even before Fidonet.

The late 1980s and early 1990s were a time of transition from the BBS to the public Internet. One BBS user described this transition as mostly seamless, with the Internet supplanting the BBS,

> For me BBSing coincides perfectly with my years in high school (1987-91). Prior to high school, I had no modem. After high school, I was in college and rarely touched my computer. But during high school, BBSing was a great way to spend the evening & not do homework ;-)
>
> My main equipment was a Commodore 64 (later Amiga 500) using a 1.2 or 2.4 kbit/second connection. Because of the slowness of the connection, everything was pure ASCII text. (There were a few graphical sites, but they would take two or three minutes to load.)
>
> In my hometown of Lancaster PA we had more BBSes than I can remember. It seemed everyone with a computer wanted to take a turn at creating a BBS, discovered it was too much work, and then pulled it down...

[63] Bush, *FidoNet: Technology, Use, Tools, and History*.

> In 1994 after having left my computer lay stagnant for a few years, I discovered something called "Mosaic" that connected to a strange entity called the world wide web. The web provided new experiences that BBSes could not provide. Places like midwinter.com or scifi.com or vidiot.com allowed me to read, learn, and find TV-related news & schedules for all my favorite shows. Plus, the web still had all my favorite downloadable files (music and swimsuit covers), plus there was the ever-reliable Usenet heirarchy where we could discuss such world-changing topics as "Who is the better captain? Picard or Sisko?"
>
> And ultimately that's what brought down BBSes. The world wide web still supplied the files, and the community atmosphere of forums, but it ALSO had vast amounts of information from [n]ational services so you could find out what's on TV, or buy books online, or whatever. The world wide web was essentially BBS 2.0. It had all the features of a local BBS, plus lots of extra bonuses only available via national providers. As such, there was no longer any need for me to continue BBSing, since the Mosaic/Netscape/Firefox browsers let me go directly to the national source.[64]

With the transition from the telephone network to the BBS, and from the BBS to the Internet, the use-case of the technology had changed, and this allowed for an expansion in the world of possibilities for hacking. This transition, as highlighted in the above passage, was seen as a reasonable progression from one type of networking technology to a slightly better version that does the same thing. The phone phreaks that were swept up in the previous transition from phones to BBSes felt something similar. It is important to note however, that these transitions were not seamless, though it may seem as if they were. Not all phone phreaks became BBS users, not

[64]Scott, *A Lancaster Teen Remembers BBSes (February 6, 2008)*.

all BBS users were hackers. Each advance in technology swept in more and more users; the barriers to entry were lowered at each step. Exploiting vulnerabilities in the telephone network in the 1960s was much harder, and required a great deal more effort and background knowledge than did dialing up to a BBS in the 1980s. Nevertheless, phreaks and hackers had a lot in common. Bruce Sterling explored the murky relationship between the two groups,

> Because the phone network pre-dates the computer network, the scofflaws known as "phone phreaks" pre-date the scofflaws known as "computer hackers." In practice, today, the line between "phreaking" and "hacking" is very blurred, just as the distinction between telephones and computers has blurred. The phone system has been digitized, and computers have learned to "talk" over phone-lines...
>
> Despite the blurring, one can still draw a few useful behavioral distinctions between "phreaks" and "hackers." Hackers are intensely interested in the "system" per se, and enjoy relating to machines. "Phreaks" are more social, manipulating the system in a rough-and-ready fashion in order to get through to other human beings, fast, cheap and under the table.[65]

According to Sterling, one of the defining characteristics of the phone phreaks, in distinction to the computer hackers, was their interest in the strictly social element of their craft above and beyond the technological element. With the introduction of computers, a small piece of the social, which had been a major part of phreaking, was lost as the technology allowed for the transition from the telephone to the computer, and from voice to text. While the seemingly most important elements of the hacker mindset had been preserved, the movement became ever so slightly less social.

[65]Sterling, *The Hacker Crackdown*.

I propose that as the technological complexity surrounding hacking grows, the human element is diminished, and the product becomes less social; it is less about human connection stemming from the lifeworld following methodologies emerging from the standpoint of life and more about instrumentality and rationalization. As hacking moves toward the far end of this spectrum, automated decision making apart from the relevant social context surrounding those decisions begins to seem more reasonable. It is only by reifying hacking and stripping it of its social roots, that automated decision making as an output of the hacking methodology can even make sense.

At the same time, these new technologies did allow for new ways for people to connect with one another, and explore the social world and their relationship with it. The transition from voice to text allowed for the creation of safe public virtual social spaces for people who were marginalized in physical social spaces. In 2003 a non-binary former BBS user gave an interview to Jason Scott about their experience on BBSes in the early 90s, describing this virtual space as the first place where they felt safe to express and live their authentic gender,

> "I got back on – I moved out in '89 and I got back on in '92 and that's when I did most of the BBSing and essentially grew myself up. I really just don't know how to phrase it, you know. BBSes really, you know, you could really completely reinvent yourself – and not on a surface level. I changed it – I changed who I was. Having – being able to be at the keys gave me more reaction time. I learned to stand up for myself – before I, you know, I was a picked-on kid and I just couldn't stand up for myself and I just didn't have enough time to think of a good response, and there was always the fear of being hit, and on the boards I completely learned how to stand up for myself, and when I had to start meeting people in person I could start saying — 'okay, what would I say online?' and just

> say that. And that worked out. Also you could be yourself without being judged in person based on your physical appearance, so I didn't have to be female. And when I was first on, male was the default, so you know, you could go with that. My — one of my first handles was gender neutral, or one of my first handles on the chat BBS — in '92 the chat BBSes were coming on — and my handle was gender neutral so that whenever someone asked me my sex, I could say, 'look at the handle,' and if they still kept asking me what sex I was, I could say, 'hey stupid, look at my handle,' and could flame them, which was always great on a lousy day if somebody earned it. Well I think that gender and sex are very often — well pretty much are always – linked together, but I think that gender is on a sliding scale, and that somebody, when they see my physical body, they draw conclusions about me that are not correct. You know they're gonna be somewhat off, but if they meet my mind first, then they have a much better idea of who I am than seeing my body, and saying 'well you're female so perhaps you wear dresses, or why don't you paint your nails, or you must like to shop for clothing' or whatever it is that girls do. I'm definitely female, but that's not the same as being a girl, and if you get to know my mind then you know who I am and you won't make any of these very odd assumptions."[66]

The ability to transform, reinvent, and reimagine the possibilities of a new technology allowed for something much more important: the transformation, reinvention, and reimagining of oneself in light of that technology or through its use. This again, is the essence of hacking, and again, the methodology is applied directly to the problems of life, and to the social world. The desire to transform the technology emerges out of the necessity of transforming the social world in response to the problems

[66]Scott, *BBS Documentary Interview: Jayne*.

of life experienced and articulated subjectively in light of one's situated knowledge. Community building is problem solving. While the first BBS users weren't doing anything quite so dramatic as breaking into the computer systems of Los Alamos national lab like The "414 Gang" of hackers in 1982,[67] they were creatively solving problems using technology, and that is a necessary condition for hacking.

The end of the BBS era and the beginning of the public Internet in the early 1990s was a time of transition, not only in terms of technology, but in terms of the meaning of hacking. The definition provided in figure 3.2 describes activities engaged in by multiple distinct groups in the early 90s, though there was significantly less of a cohesive community formed around these groups than there had been among the phone phreaks of the 60s. Once the Internet became available to the public, community cohesion among hackers decreased. While phone phreaks who met on a party line were certain to have much in common, and BBS users had, at a minimum, their interest in and knowledge about BBSes in common, the opening of the technological floodgates in the early 90s brought so many new people into virtual public spaces so quickly it was difficult for community cohesion to survive. Eric Raymond, author of The Jargon File, "a comprehensive compendium of hacker slang illuminating many aspects of hackish tradition, folklore, and humor"[68] describes the beginning of this surge in public access to previously cloistered virtual communities as the "Eternal September"[69] in reference to September 1993, when AOL first allowed its users to post on Usenet.

The hacking community in the early 1990s splintered into several sub-groups. While phone phreaking was rendered obsolete by upgrades to the telephone network, phone phreaks still existed as a community. Criminal computer hackers similar to the 414's were still active. A new type of hacker, and a new connotation of the

[67]Vollmann, *The 414s: The Original Teenage Hackers*.
[68]Raymond, *The Jargon File*.
[69]Raymond, *September that Never Ended*.

"hacker" moniker emerged at this time — those who engaged in playful, creative problem solving centered around computer programming. One example of this type of hacker is Linus Torvalds, author of the Linux Kernel, which he shared publicly on Usenet in 1991.[70] This was also a time when the business world began to recognize the revolutionary potential of the public Internet. Venture capital began to fund a wide range of business ideas that sought to leverage the Internet in order to expand markets into these new virtual spaces. This was the beginning of what would become the tech bubble of the late 1990s and eventually the emergence of Big Tech in the early 2000s. In the early 1990s, these diverse communities existed, not as a single cohesive group, but at the same time, not entirely distinct from one another either. Their cultures intermingled, and an Internet culture began to emerge. At the heart of this new culture were many of the core ideas of the hacking community: agility in response to new problems, a focus on technological innovation, and of course, creative cleverness in problem solving.

3.4 The Dominance of Big Tech

Studying the history of technology is important for understanding the present technological condition because the novelty of technological developments carries with it the deception that the collective response to a new technology is just as novel as the technology itself. The history of technology, as with all history, helps us to better understand ourselves. The understanding of the nature of the core philosophical problems associated even with new technologies can only improve through the interrogation of their placement within the broader technological and socio-historical context. The history of technology is relevant because humans have found ways of understanding, fixing, and successfully incorporating philosophically troubling technologies into society in the past, so the past can serve as a guide in this instance as

[70] Torvalds, *Linux's History*.

well.

Much of the current landscape of "Big Tech"[71] can be traced to its origins in informal technological innovation and development over more than sixty years of history, from phone phreaks in the late 1950s through the early 1980s, computer hackers in the 1980s and 1990s, the end of the tech bubble and beginning of the contemporary Silicon Valley startup culture from the late 1990s to the early 2000s, and the beginning of the dominance of the giants of Big Tech today. The history described above shows that the relationship between the individual and technology in the early days of phone phreaking and computer hacking led to a culture that prized curiosity and experimentation, that generally there were no widespread or life-threatening consequences to this approach, and in fact those who took this approach, including the founders of Apple and Microsoft, enjoyed massive financial success as a result.

This hacking culture persisted into the 1990s and gave rise to extremely important innovations like GNU/Linux, which now runs on the majority of the world's webservers (78.8% as of November 2021)[72] and supercomputers (51.6% as of November 2021).[73] In the early 2000s the same cultural attitude gave rise to Facebook, with its famous motto, "move fast and break things."[74] With each step, power consolidated into fewer and fewer hands while barriers to entry became increasingly difficult to overcome, yet the same positive attitude toward disruption persisted. The ap-

[71]I use the word "tech" here not as an abbreviation of "technology" but to describe a particular cultural milieu — centered around technology perhaps, but encompassing much more than that. The goal of this chapter is to re-situate the technological within the social and political, so when I use "tech" I am envisioning the interrelated and interdependent worlds of technology, society, politics, and economics, but in particular I am referring to a social sphere that centers itself around its relationship to technology and technological innovation while ignoring, intentionally or not, its contextual connections to the remaining areas of the social and political etc. The conglomeration of the largest, often multi-trillion dollar companies that dominate this space including Facebook, Apple, Amazon, and Google are what I am referring to when I use the term "Big Tech," rendered herein as a proper noun.

[72]World Wide Web Technology Surveys, *Usage statistics of operating systems for websites*.

[73]The Top 500 Project, *Operating System Share*.

[74]S. Ghosh, *Everything happening to Facebook stems from its radical thesis of 'Move fast and break things'*.

proach was working, and as a result, few questions were raised about its potentially problematic consequences. In its current iteration, this same approach to innovation and disruption has manifested in the form of multi-billion dollar companies developing large AI models that are likely — or in some cases known[75] — to contain serious faults that cause widespread negative impacts on populations across the world[76] yet are nevertheless used in production. Another place that this methodology appears to have enabled is the proliferation of cryptocurrencies, "programmable money" and other digital assets such as NFTs[77].

The history highlights a phenomenon that started as what can only be described as a group of bored but bright teenagers in the late 1950s and early 1960s who decided to reverse engineer the telephone system, to the contemporary culture of Silicon Valley, often derisively — though perhaps accurately — referred to as "tech-bro culture," which drives a huge portion of technology culture in the contemporary world. "Move fast and break things" is a direct descendent of this culture, which, through curiosity, seeks to learn how an existing system works in order to exploit that knowledge "for fun and profit"[78] as the saying goes. But this approach to development though innovation inevitably caused serious harm, including outright scams and frauds like Jucero[79] and Theranos[80],[81] and obvious-but-unfortunately-not-obvious-to-everyone scams like bitcoin,[82] all the way to the increasingly common surveillance technology many are willingly incorporating into our home lives such as Amazon's Alexa[83] or

[75]see for instance Bender et al. ("On the dangers of stochastic parrots: Can language models be too big")

[76]see chapter two of this dissertation

[77]Non-Fungible Tokens

[78]Tozzi and Zittrain, *For Fun and Profit: A History of the Free and Open Source Software Revolution*.

[79]Reilly, *Juicero is still the greatest example of Silicon Valley stupidity*.

[80]Carreyrou, *Bad Blood: Secrets and Lies in a Silicon Valley Startup*.

[81]Sara Randazzo, *The Elizabeth Holmes Verdict: Theranos Founder Is Guilty on Four of 11 Charges in Fraud Trial*.

[82]Griffin and Shams, "Is Bitcoin really untethered?"

[83]Falcon-Morano, *Sidewalk: The Next Frontier Of Amazon's Surveillance Infrastructure*.

Ring,[84] and other seriously problematic uses of AI, like Tesla's public beta testing of its Full Self Driving technology.[85]

Phone Phreaking in the 1960s and 1970s gave rise to a broader "hacker culture" that persisted through the remaining decades of the 20th century and, with some important modifications, became the foundation of the culture within what is now known as "Big Tech." One need not look very far for confirmation of this thesis. In the early days of their business partnership before founding Apple Computer, Steve Jobs and Steve Wozniak built and sold blue boxes. Jobs himself remarked in a 1998 interview that "if we hadn't have made blue boxes there would have been no Apple."[86] That a relationship between phone phreaking and Big Tech exists is difficult to deny, but articulating the nature of the relationship between these two worlds is of particular historical and philosophical significance.

This story began with the strange confluence of coincidences surrounding the relationship between people and the dominant social technology of the day, the telephone. The motivation to build the first blue box was arrived at almost accidentally. Phone phreaks had learned that whistling a certain tone allowed them to control the phone system, but it took more rigorous work before it would be possible to construct the first blue box. This was made possible when the first phone phreaks discovered the November 1960 volume of The Bell System Technical Journal — a private internal scientific journal published by Bell Labs. This strange relationship between corporate science and independent hackers came full circle in the early 21st century when the hacker culture inspired at least in part by these early phone phreaks helped to motivate thousands of young programmers to work on machine learning and AI and to eventually author scientific papers while employed by the likes of Facebook and

[84]Guariglia, *What to Know Before You Buy or Install Your Amazon Ring Camera*.

[85]Levin, *From swerving into a median to narrowly missing poles, videos of Tesla's latest Full Self-Driving update don't inspire much confidence*.

[86]Silicon Valley Historical Association, *Steve Jobs Interview about the Blue Box Story*.

Google[87]. The history shows that there has always been a tight relationship between individual hackers, corporate interests in the world of technology, and science. This relationship is at times antagonistic, and at others synergistic, but its existence is certain, and has been since at least the early 1960s.

What I have demonstrated so far is that the "spirit of playful cleverness,"[88] or perhaps more appropriately, a spirit of playfulness and curiosity, that characterized phreaking and hacking culture from the 1960s through the 1990s has been a persistent and defining characteristic of not only the hacking subculture, but also a core reason behind many of the technological innovations of the era. The upgrade from in-band to out-of-band signaling[89] in the telephone system was a direct result of the widespread exploitation of the vulnerabilities discovered in that system by early phone phreaks. The popularity of the BBS in the 1980s demonstrated the world-changing potential of the technology that would become known as the Internet even before the invention of hypertext transfer protocol (HTTP) by Tim Berners Lee in 1989. The groundwork for a public Internet had already been laid by the time it became technologically possible, and without the widespread popularity of the BBS, it is not certain whether HTTP would have been successful at all. The creation of the Linux Kernel in 1991 and the concurrent development of the GNU Core Utilities allowed for the preconditions necessary for independent software developers to contribute to free and open source software projects that would later change the way that both Internet infrastructure and supercomputing operate.

[87]Birhane et al. (*The Values Encoded in Machine Learning Research*) documents an alarming increase in corporate funding or affiliation in the most influential academic papers in computer science. They write that "in 2008/09, 24% of these top cited papers had corporate affiliated authors, and in 2018/19 this statistic almost tripled, to 71%"(Birhane et al., *The Values Encoded in Machine Learning Research*, pp. 7–8)

[88]Stallman, *On Hacking*.

[89]the vulnerability that allowed a phreak to seize a trunk line by introducing the 2600Hz tone existed because the machinery that controlled the telephone system was integrated with the lines that carried voice data across that system. in-line signaling refers to the ability to control the telephone system over the same line that a telephone user speaks on. Out-of-band signaling utilizes a parallel data line such that command and control of the phone system takes place on a different line than the one that carries the phone user's voice.

At the core of each of these developments is the main characteristic of hacking culture: an intrepid spirit of playfulness and curiosity, and the willingness to fearlessly engage in interventions on existing technological systems for the purpose of understanding, improving, or transforming them. Without this spirit, the Internet and computing world that we know today could not have existed. This spirit was carried forward through each era of technological development from the 1960s to the present, but in each era, its character changed slightly. Perhaps the biggest change occurred during the transition from Web 1.0 to Web 2.0[90], from the static Internet to the interactive Internet, and from the era of the personal webpage to the era of social media. But this shift occurred nearly twenty years ago, and the landscape has not remained static since. What characterizes the contemporary era is the consolidation of power in the hands of fewer and fewer large companies.

While the current landscape bears little resemblance to the Internet of the early 1990s or to the BBSes that preceded it, it does in some important cases retain some of the character of the early spirit of computer hacking. The elements of playfulness, cleverness, and curiosity have in many cases been maintained, but often bounded in ways that were not the case previously. More importantly, the attitude that prioritizes asking for forgiveness rather than permission — "move fast and break things" — is very much alive. A focus on and prioritization of innovation, exploration, and experimentation over stability and slow but steady progress has always been an important element of hacking culture, as was documented in the textfiles of the 1980s and 1990s. Hacking in these days was nimble and self-directed. It was properly

[90]The Term "Web 2.0" was coined by Darcy DiNucci in 1999(DiNucci, *Fragmented future*). Today this term is most commonly used by proponents of cryptocurrency in order to describe a possible future for the Internet that they call "Web 3.0." Given this odd, contingent relationship to the world of cryptocurrency, widely recognized to be rife with various scams and ponzi schemes, the use of the terms "Web 1.0" or "Web 2.0," for some, may imply something about my beliefs about cryptocurrency. I am not a proponent of cryptocurrency and it appears to be far too early to refer to anything as "Web 3.0." Nevertheless, the distinction between the era of an Internet marked by static webpages and one marked by interactive content is useful, and I am thus choosing to use the terms that are already in widespread use to describe this delineation.

decentralized. This attitude which puts playful curiosity and exploration first and consequences second remained in place, even as the scale of the Internet expanded such that it could no longer be nimble, self-directed, or decentralized, but instead came under the control of large, centralized institutions. In previous eras when the attitude was "consequences be damned," it wasn't a major social problem, because the consequences, positive or negative, were localized and relatively small in scope. But as the Internet scaled up and shifted to Web 2.0 in the early to mid 2000s, this approach started to produce more widespread negative consequences that have only recently become apparent.

Equally important is what changed during this time. One major aspect of hacking culture that changed during the 90s and early 2000s was the countercultural element of hacking. In the 70s and 80s hacking was done *in opposition to* big institutional players like "Ma Bell,"[91] and this is quite apparent in reading through the early textfiles of the 1980s. Jason Scott, it is worth noting, also compared these textfiles to the work of Abbie Hoffman.[92] While there was certainly interaction between the hacking world and the corporate world, this relationship was largely antagonistic.[93] In the late 1990s and early 2000s, the hacking counterculture was taken up and appropriated by big institutional powers. What used to be nothing more than some teenagers playing around with the phone in their garage became a skill that was desirable to capital. Once hacking ceased to exist in opposition to large institutional interests and started to serve these interests, the game changed permanently. People who would have otherwise been considered part of the hacker counterculture stopped war dialing[94] and

[91] After the 1984 breakup of the Bell Telephone company into multiple regional companies (e.g. Bell Atlantic, Bell South etc.) called "Baby Bells," phone phreaks referred to Bell Telephone as "Ma Bell" to mean the original whole company before the breakup.

[92] Scott, *The Text of my 1999 DEFCON 7 Speech*.

[93] For instance, many of the early BBS textfiles (philes) about the phone company cast the phone company as a villain to be defeated, an annoyance to be overcome, or occasionally, as just plain evil.

[94] War dialing, common in the early 1980s, was the process of using a computer and modem to dial a long list of numbers while a computer program notes which of those numbers were answered by another modem. The purpose of war dialing was to locate computers that could later be remotely accessed. War dialing was shown in (and takes its name from) a scene of the 1983 movie *War Games*

started learning Python and PHP; they started applying their efforts, not to playful cleverness, but to doing the things that served the interests of the big institutional players in the field. In short, hackers sold out, and the countercultural element of the colloquial understanding of the meaning of the term "hacking" was diminished. "Hacking" is now commonly taken to be synonomous with "computer programming." This certainly wasn't the case as recently as 1993 when the first Defcon[95] conference took place, or 1994 at the first HOPE[96] conference, but since the beginning of Web 2.0 and the rise of social media conglomerates like Meta (previously Facebook), many of the best hackers in the world simply opt to work for the Big Tech companies rather than for themselves. As employees of these companies, the problems to which hackers turn their attention are the problems determined by the monetary interests of multinational corporations rather than the problems which emerge from life, presenting themselves through lived experience. Even when independent, institutionally unaffiliated hackers ply their trade to, e.g., discover a new zero day exploit, there generally isn't much to do with it that doesn't put the hacker in danger of a multi-decade prison sentence except for simply turning the vulnerability in to the affected institution for a bug bounty.[97] Hacking, in this context, often lacks the imagination that it once not only had, but required. This is a direct consequence of the shift from hacking as a response to the problems of life to hacking as a service in the interests of capital.

While there has, since the 1960s, been a relationship between independent hackers and large technology companies, the nature of this relationship has changed. Once presented as a David and Goliath story with hackers painting themselves as working against the big corporate interests, each came to represent the same interests over

starring Matthew Broderick and Ally Sheedy.

[95]The largest conference for hackers and information security professionals in the world. Held every summer in Las Vegas and routinely drawing as many as 30,000 attendees

[96]Hackers on Planet Earth (HOPE) is a popular biennial hacker conference sponsored by *2600 Magazine*

[97]"Bug bounty" programs offered by many companies pay hackers who discover new vulnerabilities to disclose those them to the affected company rather than exploiting them.

time. What is interesting about this shift is that many of the elements of hacker culture which allowed for its success — especially agility and quick responsiveness to change — arose out of the counterculture and have simultaneously remained a core part of the hacker mentality even as the culture became largely subsumed under Big Tech. It was unambiguously a result of the antagonism between the phone company and the phone phreaks that such rapid progress was made in phreaking, spurred both by the thrill of playing the cat and mouse game and also a desire not to wind up in deep legal trouble, hackers developed methods not only of defeating the phone company's security, but also of covering their tracks. It truly is remarkable that after the breakup of Bell and the rise of the prominent tech companies of today, that this antagonistic relationship between hackers and corporations has been turned on its head. In the days of phone phreaking, it was only when a phreak was caught and facing legal jeopardy that they would go to work for the phone company[98] , but in the 21st century, the mark of a successful hacker is often a job offer from a FAANG[99] company.

The existence of Bell Labs during the era of phone phreaking as a counterexample to the thesis outlined above may initially stand out to readers as problematic for my argument, but I believe that the Bell Labs of the 1960s and 1970s shares little resemblance to the Googles and Facebooks of today. There are major differences between companies like Google and Facebook on the one hand and Bell Labs on

[98]After his first arrest, Joe Engressia recognized the potential to use his skills to work for the phone company, so he intentionally got caught blue boxing in order to negotiate for a job(Lapsley, *Exploding the Phone: The Untold Story of the Teenagers and Outlaws who Hacked Ma Bell*, p. 130);Charlie Pyne, Tony Lauck, and Ed Ross, three of the Harvard students described in the 1966 *Crimson* article, wrote detailed reports of their knowledge of the phone system for AT&T after their arrest by the FBI in 1963(Lapsley, *Exploding the Phone: The Untold Story of the Teenagers and Outlaws who Hacked Ma Bell*, p. 83); After his release from prison, Kevin Mitnick shared information with Sprint to help patch an open vulnerability in their system(Mitnick, Simon, and S. Wozniak, *Ghost in the Wires: My Adventures as the World's Most Wanted Hacker*, pp. 389–390)Neal Patrick, one of the members of the infamous '414 gang' of hackers who broke into the computers of the Los Alamos nuclear research facility received immunity from prosecution in exchange for sharing information about how he accomplished the hack(Vollmann, *The 414s: The Original Teenage Hackers*)

[99]Facebook, Apple, Amazon, Netflix, Google

the other, and the core of these differences has to do with their attitude toward innovation and risk. I have argued that it was the risk-taking spurred by curiosity that made hacking what it is, and it was the incorporation of this methodology into the corporate world that made 21st century Big Tech what it is. The main contribution of hackers to the world of corporate innovation was to change the risk profile. Bell Labs, though it was responsible for multiple world-changing innovations, was still, relatively speaking, a conservative company by today's tech business standards. For instance, in his history of Bell Labs, Jon Gertner offers his own comparison between the type of innovation seen at Facebook and that of the 1960s era Bell Labs,

> "One can only speculate about how [Mervin Joseph] Kelly, [John Robinson] Pierce, [William Oliver] Baker, and the rest would react to the most acclaimed American innovations of recent years — iPhones, say, or Google searches or Facebook. They would likely see them as vital, sophisticated tools for the information age. A more provocative question, however, is whether they would perceive them as paths to the future, as many economic commentators often do. Regrettably, the language that describes innovations often fails to distinguish between an innovative consumer product and an innovation that represents a leap in human knowledge and a new foundation (or 'platform,' as it is often described) for industry. In an effort to explain his motivations, Pierce once wrote in a memo, 'Things should be done only when there is the possibility of a *substantial* gain, and this must be weighed against risk.' The italics were Pierce's own."[100]

This position, one that takes risk seriously, and soberly weighs the potential transformatory power of an innovation before embarking on a project to see it through, is

[100]Gertner, *The Idea Factory: Bell Labs and the Great Age of American Innovation*, p. 344.

one that stands in stark contrast to the "move fast and break things" mentality. Similarly, the important distinction between innovations that advance human knowledge and innovations in consumer products is crucial here. The invention of the vacuum tube or the transistor, both innovations of Bell Labs, must certainly be part of the former class while virtual reality headsets or AI-enhanced targeted advertising clearly belong to the latter. The risk profile is different because under the leadership of Big Tech, the systemic risk is significantly higher[101] while the potential for substantial advances in human knowledge is considerably lower.

The more conservative views toward technological innovation expressed in the above quote by John Pierce do not hinder the development of transformatory advances in human knowledge, but they do preserve an important measure of stability in the social sphere. This attitude was expressed in the 1960s when hacking as both a methodology and a cultural force was just getting started. By the time of the foundation of Facebook, hacking was firmly planted in the cultural imagination, and the methods employed by hackers had already come to be seen as preferable to the slow and steady (though still extremely powerful) style of innovation seen in the past. The methods taken up by hackers, described in section 3.2 had, by the early 2000s, come to be seen as beneficial not just to individual problem solving, but to larger institutions as well. By taking on more risk, there was a perception that innovation would proceed more easily, but what ultimately happened was a sustained decline in social stability at the same time as an increase only in consumer innovation and not in the type of innovation that leads to substantial increases in human knowledge. It was because of their incorporation of the methodologies employed by the hackers that these consumer innovations were possible, and this of course led to major increases in profitability as a company, but the social destabilization that came as an associated cost hardly seems worth it. The major difference between the more conservative

[101] LaFrance, *The Facebook Papers: History Will Not Judge Us Kindly*.

technological innovation of the past and the more freewheeling technological innovation of the present is not chiefly the difference between a focus on human knowledge production vs. consumer product development, but of the methodological shift that underlies this ideological shift, the subsumption of "hacking" under the corporate institutional apparatus.

The cultural roots of the world of the Big Tech companies are firmly planted in the world of phone phreaking. A major difference between the two worlds is the orientation of the actors in each of these two very different spheres to the problems they are attempting to solve. The early phone phreaks were intervening on technological systems based entirely on their lived experience. It was their situatedness within this experience which drove their desire to engage with these systems. It was the context of this lived experience, not the technology itself, which drove the technological innovation spurred by the actions of these early hackers. The major difference between the early days of hacking and the current state of affairs is that the relevance of and sensitivity to the context of lived experience has been stripped away in the current iteration. The other elements of hacking as defined earlier in this chapter remain a key focus of companies like Facebook (now Meta), but their orientation toward the real problems of life has been changed by their large institutional structure, and by their obligation to return value to their shareholders. These changes flip the script of early hacking, which was motivated by the problems of life to make interventions on existing technical systems thus driving the future of innovation. Under Big Tech, the motivation for the same types of interventions is no longer the problems of life, but the desires of capital. As a publicly traded company, the primary aim and the primary driving force behind all decisions must be returning value to shareholders. Under these circumstances, even when combined with the use of the hacking methodology, the real problems of life are no longer relevant. This causes Big Tech to start with the technological innovation first and then find — or invent — the problem that it solves.

This is how Big Tech can appropriate hacking, as broadly construed as it was earlier in this chapter, and yet still end up not doing anything like the early hackers. An element of the definition of hacking that was left out, and turns out to be crucially important to its understanding, is that what motivates hackers is a sensitivity to the problems of life, and to the lifeworld context from which they emerge. What was lost along the way from the early days of hacking to the present was this recognition and understanding of the importance of context.

Given the problems associated with the subsumption of hacking under the umbrella of Big Tech, I propose the following, modified definition of hacking, which precludes the type of corporate-sponsored hacking emerging directly from the needs of capital rather than the standpoint of life.

Hacking is a form of individualistic, applied problem solving, generally centered around technology, and characterized by
(a) "playfulness, cleverness, and exploration." [a]
(b) social context sensitivity: hacking applies its energy to problems emerging from lived experience.
(c) defiance: hacking is often resistant to coercive or unjust hierarchies; it often ignores the law when it conflicts with a solution or the process for generating one.
(d) adversariality and independence: hacking is often self-defined in distinction to mainstream or orthodox systems of problem solving.
(e) allopoietic solution generation: the solutions generated through hacking are portable; they do not require others who employ those same solutions in different contexts to conform to the above characterizations a-d.

Figure 3.2: A final definition of hacking

[a] Stallman, *On Hacking*.

The addition of (d) adversariality and (b) social context sensitivity, would appear to rule out the classification of anything done under the aegis of Facebook or Google as proper "hacking." Additionally, it should not be seen as a problem that what appears to be happening at these corporations resembles hacking, as the solutions

derived through hacking are portable, and can be readily employed elsewhere. That a given institution commonly tends to take up these types of solutions and apply them towards its own set of problems does not indicate that what the institution is doing is *hacking*, but rather it is benefiting from the fruits of hacking without engaging in any hacking itself. In other words, Big Tech has merely appropriated hacking and hacking culture without ever engaging in it.

The seemingly disparate problems centered around bias in large AI models, the potential for economic destabilization from cryptocurrencies, and the politically destabilizing effects of social media are actually all part of the same longstanding phenomenon — the gradual appropriation of individual hacking by, and eventual subsumption under, the control of multi-trillion dollar technology corporations. Watching how that phenomenon has developed and changed over the past 65 or so years helps to inform us of its potential future trajectory. My goal is to articulate a social problem that has been simmering in the background of the social world for more than half a century. The problem is that, within complex technological systems, the set of successful behaviors and practices of those who productively engage with the system are different depending on the size and age of the system, its level of integration into broader society, and the shifting social context around the system. The behaviors and practices that led to successful growth of helpful new technologies in the early days of the technology are different from those that enhance systemic stability in later days after substantial growth. In exploring and analyzing the history of the concept of hacking over roughly the past sixty-five years, I have identified a central theme in engagement with cutting edge technology by skilled and technically adept "hackers." This group places an emphasis on curiosity, exploration, innovation, and fun while de-emphasizing systemic stability, and negative personal, social, and legal consequences. This set of behaviors and practices is functional for new technological systems in their early days, but becomes dysfunctional as those systems become more

integrated into the broader social sphere and as they continue to accumulate power. The reason is that as technical systems grow more complex, their distance from the problems of life grows as well; being less grounded in life, the sensitivity to this important context of those engaged with these systems begins to fade. Once sensitivity to the problems of life has waned, the activity can no longer be considered hacking. I will call activities which superficially resemble hacking but lack the necessary sensitivity to social context or other critical elements "pseudo-hacking." Pseudo-hacking has come to replace hacking in the popular imagination as the thing that hackers are engaged in. While proper hacking does still exist, pseudo-hacking predominates.

This problem, for most people, has been easy to ignore. It is hidden because of its obviousness, and as such it has been left unaddressed. The role of philosophy, broadly speaking, is to foreground these types of problems, to articulate them, to question the assumptions that underlie them, to describe their contours, and in some cases suggest alternative ways of thinking about them. The project of this chapter is one which seeks to foreground a very specific type of technological problem. One that didn't begin as a problem but as a solution to a problem. Because it provided utility for two or more generations, we became accustomed to ignoring it and assuming that it was something that it no longer is and possibly never was. The work of articulation of this particular problem is philosophical because it is grounded in an attempt to draw this problem out into the light so that it can be confronted. Without philosophy, we cannot expect to understand why Big Tech is engaged so heavily in tech-solutionism, why it continues to make impossible-to-fulfill promises, why it struggles to become more inclusive, equitable, and diverse, how all of these problems are situated within a context of late capitalism, and how that context reinforces the problem and short circuits solutions. In exploring this issue in its full context, I hope to illustrate how deep its roots go in order to define the scope of the task ahead as we attempt, together, to steer this ship in a new direction. This has always been the role of philosophy in

a tradition going back to Socrates.

Engaging in a set of behaviors and practices in one context and then repeating them in a radically different context can result in so many qualitative changes that it's hard to say that the behaviors and practices really were the same. The early phone phreaks were curious, technically adept, excited about exploring something new, and didn't think much about consequences. That's the same thing that happened with many BBS users in the 80s[102], the early 1990s free and open source software pioneers like Linus Torvalds, and the successful early 2000s startups including Facebook, now Meta. The process, mentality, and spirit is the same — each is engaged in something that resembles hacking. What's different is the changing context around what they're doing that makes what they're doing qualitatively and quantitatively different.

3.5 Conclusion

While it is clear that the methods and practices employed by those who today call themselves hackers share, in some important measure, a historical lineage traceable back to the phone phreaks of the late 1960s and early 1970s, it is also clear that many of the characteristics of hacking described above in Figure 3.4 did not survive into the present. (a) is certainly still a driving force, as it really is the essence of hacking. It is likely that (c) has survived as well. But each of these other components of hacking have faded away as both the share of institutional power and the number of external financial incentives in this space grew, and as the relationship between the individual hacker and the institution became less antagonistic and more cooperative. Of those past characteristics of hacking that have been lost, the most consequential is (b): recognition of the social context around technology and an orientation toward solving

[102]for example Jason Scott of Textfiles.com who has thoroughly recorded his experience of those early BBS days and archived many of the posts made on those boards. Scott was too late for proper phone phreaking, but had be been born ten years earlier, he may well have encountered Joe Engressia and John Draper on a party line

the problems of life, subjectively determined out of one's situatedness. Without this element, whatever else an activity is, it isn't hacking.

This shift mirrors an ongoing rift in the public, non-academic, understanding of epistemology. The unfortunately widespread view of epistemology remains one that believes science is in the business of getting at and accumulating objective Truths. While this view has been on the decline within academic circles since Kuhn,[103] it is still very much alive in the popular imagination among non-scientists, among whom are futurists, tech innovators, tech business leaders, tech influencers, and so on. While the early hackers did not necessarily reject this view explicitly and openly embrace some form of social epistemology, that is what they were doing implicitly from the start. One of the biggest markers of the shift from the early hackers to the world of Big Tech is the shift from social epistemology to the resurgence of the View From Nowhere.

In today's tech world, there is, too often, no acknowledgement that the latest technological innovation is situated within a broader context that includes a system of social relations, and that acting on that system will have consequences that will ripple through it. The trajectory of this particular orientation to technology over the past twenty or so years has been marked by a stubborn refusal to acknowledge the inter-relatedness of the social and technological worlds[104]. There has been a focus on tech-solutionism — uncritically applying technological solutions to explicitly social problems — rather than the integration of the technological and social as was common in the early days of hacking. As the Big Tech companies grew larger, the problems became more entrenched. The end of this historical narrative is a set of new technologies that are explicitly built to be divorced from the human socio-historical context. They're disruptive for the sake of being disruptive, and they aim to solve problems that largely don't exist. They create problems no one anticipated because

[103]Kuhn, *The Structure of Scientific Revolutions: 50th Anniversary Edition*.

[104]In fact it may not be a matter of inter-relatedness as the distinction itself is questionable.

no one took the time to think through the potential consequences.

The error of tech-solutionism is the desire to strip something of its social context and of its situatedness in order to allow the supposedly objective technology to intervene upon it. Objectivity is not diminished by the presence of more context around an issue at the heart of a pending intervention, in fact, it is enhanced by it. According to Donna Haraway, "Feminist objectivity means quite simply situated knowledges."[105] The resistance to thinking about and understanding context within the tech space is a self-reinforcing problem that quickly spirals out of control. This is one of the distinguishing markers between the earlier forms of hacking, based as they were in respect for situated knowledge, and the new form of pseudo-hacking under Big Tech, which carries this false sense objectivity. The tech-solutionist epistemology now favored in the mythology of Silicon Valley stands in stark contrast to the situated knowledges of the early hackers. The fatal error of tech-solutionism is the belief in the possibility of the view from nowhere, or as Haraway describes it, the god trick. The production and reproduction of god trick "techno-monsters"[106] has now become the primary output of the pseudo-hacking methodology used by Big Tech, and they exist not only in artificial vision, but in all corners of the information technology ecosystem. The impulse is to solve epistemology through artificial omnipotence, but the result has been the obliteration of any critical voices who have spoken out against the futility and harm of such a project. The stories and perspectives of those who have been marginalized through this process have often been lost to history, and the erasure of these situated knowledges has thereby harmed the objectivity of the entire information technology space.

A similar point is made by Melissa Villa-Nicholas in her excellent history of Latinas as information and telecommunication workers. Villa-Nicholas criticizes the erasure of

[105]Haraway, "Situated knowledges: The science question in feminism and the privilege of partial perspective", p. 581.

[106]Haraway, "Situated knowledges: The science question in feminism and the privilege of partial perspective", p. 581.

these workers as part of the continued illusion of the neutrality of telecommunications technology, "the invisibility of Latina information labor promulgates the ideology that technology is neutral and uninfluenced by the relations of powers that dictate everyday life."[107] Recovering and telling this history is important because it acts as a counterweight to the common myth that tech billionaires are self-made. On the contrary, Villa-Nicholas argues, tech CEOs are "powered by many people working at different levels of information behavior."[108]

To recover this hidden history is to recognize the lost potential for objectivity according to the standards of objectivity in science established by feminist and social epistemologists. Helen Longino argues that "a method of inquiry is objective to the degree that it permits transformative criticism."[109] One key to objectivity is the capacity of a community to engage in a critical dialogue. What becomes apparent in the shift from early hacking to pseudo-hacking under Big Tech is that this critical dialogue, in the past, was critical of corporate power, and has today been subsumed under that very power. The capacity of the scientific community to criticize core structures of power under Big Tech, the seat of power with regard to scientific publications in the field of AI, has recently been credibly called into question. In late 2020, several Google employees authored a paper critical of the environmental harms caused by the training of large language models by Google,[110] and the authors of the paper were fired by Google in retaliation. The degree to which objectivity can be assumed in an environment where 71% of the top-cited computer science papers have one or more corporate-affiliated authors[111] must be questioned.

The trajectory of the transition from the situated knowledges of the early hackers

[107] Villa-Nicholas, *Latinas on the Line: Invisible Information Workers in Telecommunications*, p. 5.
[108] Villa-Nicholas, *Latinas on the Line: Invisible Information Workers in Telecommunications*, p. 5.
[109] Longino, *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*, p. 76.
[110] Bender et al., "On the dangers of stochastic parrots: Can language models be too big".
[111] Birhane et al., *The Values Encoded in Machine Learning Research*, p. 8.

to the consolidated power of pseudo-hacking under Big Tech demonstrates that the clear epistemic trend in tech has been away from sensitivity to the social context and toward a decontexualized, false objectivity. The impulse behind this transition has been the consolidation of power, and its primary effect has been to cut off critical conversations and shut out critical voices.

This history of hacking described in this chapter shows that the path of technological innovation through the methodology employed by hackers focuses by necessity on technological rather than social systems. The experimentation and exploration of these systems does engage with the social world, but when rendered as pseudo-hacking, does not recognize it as an integral part of the system. The relevant system that can be experimented with is the technological one. While there is a robust use of social engineering as a tactic in hacker history, it is now used as a means to an end. Humans are seen as gatekeepers to the system, not necessarily as part of it. This approach also discounts the relevance of human users of the system *as a part of the system*. Collectively, the other users of the telephone system, the BBS, the Internet, the social network, the Bitcoin network, the AI algorithm and so on, are as much a part of those systems as the technological components that the hackers are interested in exploring and understanding. In other words, the trajectory of pseudo-hacking, broadly construed, is not interested in the sociality of these systems except to the extent that understanding social factors is a precondition for engaging with the technology. This is, in fact, what generally must be discounted before hacking a system — the phone phreaks knew their actions were prohibited by both the phone company and the law. The late 1980s and early 1990s computer hackers often accessed systems without the consent of the systems owner, and did so knowingly. The legal and social norms that exist in "meatspace" simply do not translate into virtual spaces for this particular group. Those norms, mores and even laws are often ignored. The methodology of the hacker, as I have argued, has been appropriated by and incorporated

into the ethos of many of the largest tech companies in the world today — companies which own and operate virtual spaces used by billions of people on a daily basis. The consequences of the combination of a disregard for the law and for the consequences of one's actions with the absolute power associated with literal ownership of all extant public virtual spaces are potentially disastrous. One of the defining features of the pseudo-hacking methodology is a focus on technology as distinct and set apart from people and from the social world. It's no wonder then that tech-solutionism should thrive under these conditions and we should see so many technological solutions to social problems.

The better solution is to recognize that at no time has the technology been truly separate from the social or political spheres. It has always been integrated into one techno-socio-political system. To engage with only one element of this system and deny its integration with the others is dangerous precisely because of the inextricability and interdependence of the parts of the system. Any intervention on the technological element will have effects on the social or political elements. The trick the pseudo-hacker methodology plays is in allowing proponents of the methodology to believe that their interventions will be limited to technology only. This is not, and cannot be the case. Intervening on an existing technology in use by people will have effects on people, and should thus be treated as an intervention directly on people. To "move fast and break things" will undeniably mean moving fast and breaking people. In the 21st century we must be more deliberate in our interventions on technical systems than the hackers of the 1960s - 1980s had the freedom to be. We must recognize, categorize, and enumerate the potential social harms of a technological intervention *before* deploying it. The time for moving fast and breaking things is over. It is time to move slow and respect people.

Bibliography

2600 Magazine. *Hacker Perspective Submissions*. https://twitter.com/2600/status/978375671566266368. Accessed: 2022-02-07. 2018.

60 Minutes Australia. *How con-artist Anna Sorokin ripped off the New York elite and became a star*. https://www.youtube.com/watch?v=GQbNnUW_xqw. Accessed: 2022-02-26. 2021.

Abdalla, Mohamed and Moustafa Abdalla. *The Grey Hoodie Project: Big Tobacco, Big Tech, and the threat on academic integrity*. 2021. arXiv: 2009.13676 [cs.CY].

Ahmed, Nur and Muntasir Wahed. *The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research*. 2020. arXiv: 2010.15581 [cs.CY].

Akbulut, Y., A. Şengür, and S. Ekici. "Gender recognition from face images with deep learning". In: *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*. 2017, pp. 1–4. DOI: 10.1109/IDAP.2017.8090181.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *Proceedings of ICLR*. 2015.

Barbrook, Richard and Andy Cameron. "The californian ideology". In: *Science as Culture* 6.1 (1996), pp. 44–72.

Batterman, Robert W. and Collin C. Rice. "Minimal Model Explanations". In: *Philosophy of Science* 81.3 (2014), pp. 349–376. DOI: 10.1086/676677. eprint: https://doi.org/10.1086/676677. URL: https://doi.org/10.1086/676677.

BBC. *A Call From Joybubbles*. https://www.bbc.co.uk/programmes/b08hlnjq. Accessed: 2021-11-19. 2018.

Bender, Emily M et al. "On the dangers of stochastic parrots: Can language models be too big". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, NY, USA*. 2021.

Bennett, Cynthia L and Os Keyes. "What is the Point of Fairness? Disability, AI and The Complexity of Justice". In: *Workshop on AI Fairness for People with Disabilities at ACM SIGACCESS Conference on Computers and Accessibility*. 2019.

Bero, Lisa A. "Tobacco industry manipulation of research." In: *Public health reports* 120.2 (2005), pp. 200–208.

Bevard, Charles W. *Five Students Psych Bell System, Place Free Long Distance Calls*. https://www.thecrimson.com/article/1966/5/31/five-students-psych-bell-system-place/. Accessed: 2021-11-17. 1966.

Birhane, Abeba et al. *The Values Encoded in Machine Learning Research*. 2021. arXiv: 2106.15590 [cs.LG].

Blaise Agüera y Arcas, Margaret Mitchell and Alexander Todorov. *Physiognomy's New Clothes*. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a. Accessed: 2022-02-16. 2022.

Blankenship, Loyd. *The Conscience of a Hacker*. http://phrack.org/issues/7/3.html. Accessed: 2022-02-07. 1986.

Bokulich, Alisa. "Distinguishing Explanatory from Nonexplanatory Fictions". In: *Philosophy of Science* 79.5 (2012), pp. 725–737. DOI: 10.1086/667991. eprint: https://doi.org/10.1086/667991. URL: https://doi.org/10.1086/667991.

— "How scientific models can explain". In: *Synthese* 180.1 (2011), pp. 33–45.

— "Searching for Non-Causal Explanations in a Sea of Causes". In: *Explanation Beyond Causation*. Ed. by Alexander Reutlinger and Juha Saatsi. Oxford: Oxford University Press, 2018. Chap. 7, pp. 141–163.

Breen, C. and C. A. Dahlbom. "Signaling Systems for Control of Telephone Switching". In: 39.6 (Nov. 1960), pp. 1381–1444. ISSN: 0005-8580 (print), 2376-7154 (electronic). URL: https://archive.org/details/bstj39-6-1381.

Broussard, M. *Artificial Unintelligence: How Computers Misunderstand the World*. Mit Press. MIT Press, 2019. ISBN: 9780262537018.

Buckner, Cameron. "Deep Learning: A Philosophical Introduction". In: *Philosophy Compass* 14.10 (2019). DOI: 10.1111/phc3.12625.

Bush, Randy. *FidoNet: Technology, Use, Tools, and History*. https://www.fidonet.org/inet92_Randy_Bush.txt. Accessed: 2021-12-14. 1992.

Carreyrou, J. *Bad Blood: Secrets and Lies in a Silicon Valley Startup*. Knopf Doubleday Publishing Group, 2018. ISBN: 9781524731663.

Ceglowski, Maciej. *Privacy Rights and Data Collection in a Digital Economy*. 2019. URL: https://www.banking.senate.gov/hearings/privacy-rights-and-data-collection-in-a-digital-economy (visited on 03/10/2021).

Chen, Jiawei et al. "AutoDebias: Learning to Debias for Recommendation". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 21–30. ISBN: 9781450380379. DOI: 10.1145/3404835.3462919. URL: https://doi.org/10.1145/3404835.3462919.

Coeckelbergh, M. *AI Ethics*. The MIT Press Essential Knowledge series. MIT Press, 2020. ISBN: 9780262538190.

Constine, Josh. *Facebook changes mission statement to 'bring the world closer together'*. https://techcrunch.com/2017/06/22/bring-the-world-closer-together/. Accessed: 2021-12-2. 2017.

Crawford, K. *The Atlas of AI*. Yale University Press, 2021. ISBN: 9780300209570.

Darwiche, Adnan and Auguste Hirth. *On The Reasons Behind Decisions*. 2020. arXiv: 2002.09284 [cs.AI].

Dastin, Jeffrey. *Amazon scraps secret AI recruiting tool that showed bias against women*. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (visited on 03/16/2021).

Department of Justice. *Theranos Founder Elizabeth Holmes Found Guilty Of Investor Fraud*. https://www.justice.gov/usao-ndca/pr/theranos-founder-elizabeth-holmes-found-guilty-investor-fraud. Accessed: 2022-02-26. 2022.

— *William McFarland Sentenced To 6 Years In Prison In Manhattan Federal Court For Engaging In Multiple Fraudulent Schemes And Making False Statements To*

A Federal Law Enforcement Agent. https://www.justice.gov/usao-sdny/pr/william-mcfarland-sentenced-6-years-prison-manhattan-federal-court-engaging-multiple. Accessed: 2022-02-26. 2018.

Devlin, Jacob et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423. URL: https://doi.org/10.18653/v1/n19-1423.

DiNucci, Darcy. *Fragmented future*. http://darcyd.com/fragmented_future.pdf. Accessed: 2021-12-12. 1999.

Doorbell, Evan. *How I Became a Phone Phreak: The Dark Side of 'Party Lines,' November 1970*. http://www.evan-doorbell.com/production/HowBPhreak07-rough.mp3. Accessed: 2021-12-12. 2021.

Douglas, Heather. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press, 2009.

Draper, John. *Crunch Life 01: Denny Teresi*. https://www.youtube.com/watch?v=i2CCiI-c2qY. Accessed: 2021-12-12. 2014.

Eklund, Matti. "Fictionalism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University, 2019.

Engressia, Joe and Andrew Huse. *Joybubbles (Joe Engressia)*. https://digital.lib.usf.edu/?u23.38. Accessed: 2021-11-19. 2004.

Erlanger, O. and L.O. Govela. *Garage*. MIT Press, 2019. ISBN: 9780262347839.

Eubanks, V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group, 2018. ISBN: 9781466885967.

Falcon-Morano, Emiliano. *Sidewalk: The Next Frontier Of Amazon's Surveillance Infrastructure*. https://www.aclu.org/news/privacy-technology/sidewalk-the-next-frontier-of-amazons-surveillance-infrastructure/. Accessed: 2021-12-24. 2021.

Feenberg, Andrew. *Technosystem: The Social Life of Reason*. Harvard University Press, 2017. ISBN: 9780674971783.

— "Ten Paradoxes of Technology". In: *Techné: Research in Philosophy and Technology* 14.1 (2010), pp. 3–15.

— *Transforming Technology: A Critical Theory Revisited*. Oxford University Press, 2002. ISBN: 9780195146158.

— "Democratic Rationalization: Technology, Power, and Freedom". In: *Philosophy of Technology: The Technological Condition: An Anthology*. Ed. by Robert C. Scharff and val Dusek. Sussex, UK: Wiley-Blackwell, 2014, pp. 706–719.

Field, H. *Science without Numbers*. OUP Oxford, 2016. ISBN: 9780191083778.

Fraassen, B.C. van, Oxford University Press, and P.P.B.C. Van Fraassen. *The Scientific Image*. Clarendon Library of Logic and Philosophy. Clarendon Press, 1980. ISBN: 9780198244271.

Fuster, Andreas et al. "Predictably unequal? the effects of machine learning on credit markets". In: *The Effects of Machine Learning on Credit Markets (November 6, 2018)* (2018).

Gertner, J. *The Idea Factory: Bell Labs and the Great Age of American Innovation*. Penguin Press, 2013. ISBN: 9780143122791.

Ghosh, Shona. *Everything happening to Facebook stems from its radical thesis of 'Move fast and break things'*. https://www.businessinsider.com/everything-happening-to-facebook-stems-from-its-radical-thesis-of-move-fast-and-break-things-2018-3?op=1. Accessed: 2021-10-29. 2018.

Gilbertson, Scott. *Feb. 16, 1978: Bulletin Board Goes Electronic*. https://www.wired.com/2010/02/0216cbbs-first-bbs-bulletin-board/. Accessed: 2021-12-13. 2010.

Goldstein, E. *The Best of 2600: A Hacker Odyssey*. Wiley, 2008. ISBN: 9780470403419.

Goldstein, Emmanuel. "Ahoy!" In: *2600 Magazine* 1 (Jan. 1984), p. 1. URL: https://archive.org/details/2600magazine/2600_1-1/mode/2up.

Golumbia, D. *The Politics of Bitcoin: Software as Right-Wing Extremism*. Forerunners: Ideas First. University of Minnesota Press, 2016. ISBN: 9781452953816.

Graeber, D. *The Utopia of Rules: On Technology, Stupidity, and the Secret Joys of Bureaucracy*. Melville House, 2016. ISBN: 9781612195186.

Griffin, John M and Amin Shams. "Is Bitcoin really untethered?" In: *The Journal of Finance* 75.4 (2020), pp. 1913–1964.

Guariglia, Matthew. *What to Know Before You Buy or Install Your Amazon Ring Camera*. https://www.eff.org/deeplinks/2020/02/what-know-you-buy-or-install-your-amazon-ring-camera. Accessed: 2021-12-24. 2020.

Hall, Shannon. *Exxon Knew about Climate Change almost 40 years ago*. https://www.scientificamerican.com/article/exxon-knew-about-climate-change-almost-40-years-ago/. Accessed: 2022-03-09. 2015.

Haraway, Donna. "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century". In: *Manifestly Haraway*. Ed. by Donna Haraway. Posthumanities. Minneapolis: University of Minnesota Press, 2016. Chap. 1, pp. 5–90. ISBN: 9780816650484.

— *Primate Visions: Gender, Race, and Nature in the World of Modern Science*. New York: Routledge, 1989.

— "Situated knowledges: The science question in feminism and the privilege of partial perspective". In: *Feminist studies* 14.3 (1988), pp. 575–599.

Harrison, Brent, Upol Ehsan, and Mark O. Riedl. "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations". In: *CoRR* abs/1702.07826 (2017). arXiv: 1702.07826. URL: http://arxiv.org/abs/1702.07826.

Harvard University Library. *Scientific Racism*. https://library.harvard.edu/confronting-anti-black-racism/scientific-racism. Accessed: 2022-02-28. 2022.

Heidegger, M. and J. van Buren. *Ontology—The Hermeneutics of Facticity*. Studies in Continental Thought. Indiana University Press, 2008. ISBN: 9780253004468.

Hempel, Carl G. "The Function of General Laws in History". In: *The Journal of Philosophy* 39.2 (1942), pp. 35–48. ISSN: 0022362X. URL: http://www.jstor.org/stable/2017635.

Hempel, Carl G. and Paul Oppenheim. "Studies in the Logic of Explanation". In: *Philosophy of Science* 15.2 (1948), pp. 135–175.

Hicks, M. *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing*. History of Computing. MIT Press, 2018. ISBN: 9780262535182.

Hicks, Mar. "When Did the Fire Start?" In: *Your Computer is On Fire*. Ed. by T.S. Mullaney et al. Cambridge: The MIT Press, 2021. Chap. 9, pp. 11–26.

Hill, Kashmir. *Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match*. 2020. URL: https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html (visited on 03/10/2021).

Hilts, Philip J. *Tobacco Chiefs Say Cigarettes Aren't Addictive*. https://www.nytimes.com/1994/04/15/us/tobacco-chiefs-say-cigarettes-aren-t-addictive.html. Accessed: 2022-02-21. 1994.

Hoffman, A. and L. Fithian. *Steal This Book (50th Anniversary Edition)*. Hachette Books, 2021. ISBN: 9780306847189.

Hoffman, Abbie. *Youth International Partyline Newsletter*. https://archive.org/details/YIPLTAP_1-91. Accessed: 2021-11-22. 1971.

Hoffmann, Anna Lauren, Nicholas Proferes, and Michael Zimmer. ""Making the world more open and connected": Mark Zuckerberg and the discursive construction of Facebook and its users". In: *New Media & Society* 20.1 (2018), pp. 199–218. DOI: 10.1177/1461444816660784. eprint: https://doi.org/10.1177/1461444816660784. URL: https://doi.org/10.1177/1461444816660784.

Hogan, Teresa, Elaine Hutson, and Paul Drnevich. "Drivers of External Equity Funding in Small High-Tech Ventures". eng. In: *Journal of small business management* 55.2 (2017), pp. 236–253. ISSN: 0047-2778.

Holmes, Elizabeth. *TED MED 2014*. https://www.youtube.com/watch?v=SX7ec3uDlhs. Accessed: 2022-02-27. 2014.

Inwagen, P. van. *Material Beings*. Cornell paperbacks. Cornell University Press, 1995. ISBN: 9780801483066.

Jain, Sarthak and Byron C Wallace. "Attention is not Explanation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 3543–3556.

Jasanoff, S. *States of Knowledge: The Co-Production of Science and the Social Order*. International Library of Sociology. Taylor & Francis, 2004. ISBN: 9781134328338.

Kaiser, B. *Targeted: The Cambridge Analytica Whistleblower's Inside Story of How Big Data, Trump, and Facebook Broke Democracy and How It Can Happen Again*. Harper, 2019. ISBN: 9780062965806.

Kalluri, Pratyusha. "Don't ask if artificial intelligence is good or fair, ask how it shifts power". In: *Nature* 583.7815 (July 2020), p. 169. ISSN: 0028-0836. DOI: 10.1038/d41586-020-02003-2. URL: https://doi.org/10.1038/d41586-020-02003-2.

Kaplan, Michael. *Three Blind Phreaks*. https://www.wired.com/2004/02/phreaks-2/. Accessed: 2021-12-2. 2004.

Khademi, Aria and Vasant Honavar. "Algorithmic Bias in Recidivism Prediction: A Causal Perspective". In: *ArXiv* abs/1911.10640 (2019).

King, Taran. "Introduction..." In: *Phrack Magazine* 1 (Nov. 1985), p. 1. URL: http://www.phrack.org/issues/1/1.html.

Kitcher, P. "Explanatory Unification and the Causal Structure of the World". In: *Scientific Explanation*. Ed. by P. Kitcher and W.C. Salmon. Minneapolis: University of Minnesota Press, 1989. Chap. 9, pp. 410–505.

Kroon, F., J. McKeown-Green, and S. Brock. *A Critical Introduction to Fictionalism*. Bloomsbury Critical Introductions to Contemporary Metaphysics. Bloomsbury Publishing, 2018. ISBN: 9781472513946.

Kuhn, T.S. *The Structure of Scientific Revolutions: 50th Anniversary Edition*. University of Chicago Press, 2012. ISBN: 9780226458144.

Kumar, Sawan and Partha Talukdar. *NILE : Natural Language Inference with Faithful Natural Language Explanations*. 2020. arXiv: 2005.12116 [cs.CL].

LaFrance, Adrienne. *The Facebook Papers: History Will Not Judge Us Kindly*. https://www.theatlantic.com/ideas/archive/2021/10/facebook-papers-democracy-election-zuckerberg/620478/. Accessed: 2021-12-24. 2021.

Lapsley, Phil. *Bill Acker, 1953-2015*. https://blog.historyofphonephreaking.org/2015/09/bill-acker-1953-2015.html. Accessed: 2021-12-2. 2015.

— *Exploding the Phone: The Untold Story of the Teenagers and Outlaws who Hacked Ma Bell*. Grove Atlantic, 2013. ISBN: 9780802193759.

— *Phone Phreak Confidential: The Backstory On The History Of Phreaking*. https://www.youtube.com/watch?v=15q4m6YO1v8&list=TLPQMTkxMTIwMjGye2-atHSyfg&index=2. Accessed: 2021-11-19. 2012.

Latour, B., A. Sheridan, and J. Law. *The Pasteurization of France*. Harvard University Press, 1993. ISBN: 9780674657618.

Laudan, Larry. *Science and Values: The Aims of Science and Their Role in Scientific Debate*. University of California Press, 1984.

Lee, Timothy B. *Richard Stallman leaves MIT after controversial remarks on rape*. https://arstechnica.com/tech-policy/2019/09/richard-stallman-leaves-mit-after-controversial-remarks-on-rape/. Accessed: 2022-01-14. 2019.

Legg, Tess, Jenny Hatchard, and Anna B Gilmore. "The Science for Profit Model—How and why corporations influence science and the use of science in policy and practice". eng. In: *PloS one* 16.6 (2021), e0253272–e0253272. ISSN: 1932-6203.

Lepikhin, Dmitry et al. *GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding*. 2020. arXiv: 2006.16668 [cs.CL].

Levin, Tim. *From swerving into a median to narrowly missing poles, videos of Tesla's latest Full Self-Driving update don't inspire much confidence*. https://www.businessinsider.com/tesla-fsd-full-self-driving-videos-flaws-glitches-2021-7. Accessed: 2021-12-24. 2021.

Lichstein, Henry. *Telephone Hackers Active*. https://thetech.com/issues/83/24/pdf. Accessed: 2022-01-13. 1963.

Liquid Bug. *Screwing with School Computers*. http://textfiles.com/hacking/school.txt. Accessed: 2022-02-07. 1993.

Longino, Helen. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, 1990.

Maloney, Carolyn B. *Cracking Down on Ransomware: Strategies for Disrupting Criminal Hackers and Building Resilience Against Cyber Threats*. 2021. URL: https://oversight.house.gov/legislation/hearings/cracking-down-on-ransomware-strategies-for-disrupting-criminal-hackers-and (visited on 11/22/2021).

Marechal, Catherine et al. "Survey on AI-Based Multimodal Methods for Emotion Detection". In: *High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet*. Ed. by Joanna Kołodziej and Horacio González-Vélez. Cham: Springer International Publishing, 2019, pp. 307–324. ISBN: 978-3-030-16272-6. DOI: 10.1007/978-3-030-16272-6_11. URL: https://doi.org/10.1007/978-3-030-16272-6_11.

Mark Colyvan, John Cusbert and Kelvin McQueen. "Two Flavours of Mathematical Explanation". In: *Explanation Beyond Causation*. Ed. by Alexander Reutlinger and Juha Saatsi. Oxford: Oxford University Press, 2018. Chap. 11, pp. 231–249.

McCarthy, Cormac. *No Country for Old Men*. Vintage International. Random House, 2005. ISBN: 9780375706677.

McIlwain, C.D. *Black Software: The Internet & Racial Justice, from the AfroNet to Black Lives Matter*. Oxford University Press, 2019. ISBN: 9780190863852.

Meisenzahl, Mary. *These 5 tech companies started in garages, and now they're worth billions. These are their modest beginnings*. https://www.businessinsider.in/tech/enterprise/news/these-5-tech-companies-started-in-garages-and-now-theyre-worth-billions-these-are-their-modest-beginnings-/articleshow/72917528.cms. Accessed: 2022-02-19. 2019.

Menon, Sachit et al. *PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models*. 2020. arXiv: 2003.03808 [cs.CV].

Meta. *Our Mission*. https://about.facebook.com/company-info/. Accessed: 2021-12-3. 2021.

Metz, Cade and Daisuke Wakabayashi. *Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.* https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html. Accessed: 2022-03-18. 2020.

Mitnick, K., W.L. Simon, and S. Wozniak. *Ghost in the Wires: My Adventures as the World's Most Wanted Hacker*. Little, Brown, 2012. ISBN: 9780316037723.

Morrison, Margaret. *Reconstructing Reality: Models, Mathematics, and Simulations*. Oup Usa, 2015.

Mozur, Paul. *One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority*. 2019. URL: https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html (visited on 03/10/2021).

Murshed, M. G. Sarwar et al. *Hazard Detection in Supermarkets using Deep Learning on the Edge*. 2020. arXiv: 2003.04116 [cs.CV].

Museum, Computer History. *Timeline of Computer History*. https://www.computerhistory.org/timeline/. Accessed: 2021-12-13. 2021.

Nader, Ralph. *Unsafe at Any Speed: The designed-in dangers of the American automobile*. New York: Grossman Publishers, 1965.

Narang, Sharan et al. *WT5?! Training Text-to-Text Models to Explain their Predictions*. 2020. arXiv: 2004.14546 [cs.CL].

Ng, Choon Boon, Yong Haur Tay, and Bok Min Goi. *Vision-based Human Gender Recognition: A Survey*. 2012. arXiv: 1204.1611 [cs.CV].

Noble, S.U. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. ISBN: 9781479837243.

Noble, Safiya Umoja. "Your Robot Isn't Neutral". In: *Your Computer is On Fire*. Ed. by T.S. Mullaney et al. Cambridge: The MIT Press, 2021. Chap. 9, pp. 199–212.

Noorman, Merel. "Computing and Moral Responsibility". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2020. Metaphysics Research Lab, Stanford University, 2020.

O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016. ISBN: 9780553418828.

Pfizer. *Our Path to Developing the Pfizer-BioNTech COVID-19 Vaccine*. https://www.pfizer.com/science/coronavirus/vaccine/rapid-progress. Accessed: 2022-02-20. 2022.

Pithan, D.M. *Corporate Research Laboratories and the History of Innovation*. Management, organizations and society. Routledge, 2021. ISBN: 9780367476601.

Potochnik, A. *Idealization and the Aims of Science*. University of Chicago Press, 2017. ISBN: 9780226507194.

Potochnik, Angela. "Explanatory Independence and Epistemic Interdependence: A Case Study of the Optimality Approach". In: *The British Journal for the Philosophy of Science* 61.1 (June 2009), pp. 213–233. ISSN: 0007-0882. DOI: 10.1093/bjps/axp022. eprint: https://academic.oup.com/bjps/article-pdf/61/1/213/4260579/axp022.pdf. URL: https://doi.org/10.1093/bjps/axp022.

— "Optimality Modeling and Explanatory Generality". In: *Philosophy of Science* 74.5 (2007), pp. 680–691. DOI: 10.1086/525613. eprint: https://doi.org/10.1086/525613. URL: https://doi.org/10.1086/525613.

Poulsen, Kevin. *FBI Charges Blind Phone Phreak With Intimidating a Verizon Security Official*. https://www.wired.com/2008/06/blind-teenage-h/. Accessed: 2021-12-2. 2008.

Puri, Nikaash et al. *Explain Your Move: Understanding Agent Actions Using Specific and Relevant Feature Attribution*. 2019. arXiv: 1912.12191 [cs.CV].

Raymond, Eric. *September that Never Ended*. http://www.catb.org/jargon/html/S/September-that-never-ended.html. Accessed: 2022-02-04. 1996.

— *The Jargon File*. http://www.catb.org/jargon/html/online-preface.html. Accessed: 2022-02-04. 1991.

Reilly, Claire. *Juicero is still the greatest example of Silicon Valley stupidity*. https://www.cnet.com/news/juicero-is-still-the-greatest-example-of-silicon-valley-stupidity/. Accessed: 2021-10-29. 2018.

Reutlinger, Alexander. "Extending the Counterfactual Theory of Explanation". In: *Explanation Beyond Causation*. Ed. by Alexander Reutlinger and Juha Saatsi. Oxford: Oxford University Press, 2018. Chap. 4, pp. 74–95.

Rice, Collin. "Idealized Models, Holistic Distortions, and Universality". In: *Synthese* 195.6 (2018), pp. 2795–2819.

— "Moving Beyond Causes: Optimality Models and Scientific Explanation". In: *Nous* 49.3 (2015), pp. 589–615.

Rohwer, Yasha and Collin Rice. "How are Models and Explanations Related?" In: *Erkenntnis* 81.5 (2016), pp. 1127–1148.

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

Safra, Lou et al. "Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings". In: *Nature Communications* 11.1 (Sept. 2020), p. 4728. ISSN: 2041-1723. DOI: 10.1038/s41467-020-18566-7. URL: https://doi.org/10.1038/s41467-020-18566-7.

Salmon, W.C. *Scientific Explanation and the Causal Structure of the World*. LPE Limited Paperback Editions. Princeton University Press, 1984. ISBN: 9780691101705.

— *Statistical Explanation and Statistical Relevance*. Online access: JSTOR Books at JSTOR. University of Pittsburgh Press, 1971. ISBN: 9780822974116.

Sara Randazzo Heather Somerville, Christopher Weaver. *The Elizabeth Holmes Verdict: Theranos Founder Is Guilty on Four of 11 Charges in Fraud Trial*. https://www.wsj.com/articles/the-elizabeth-holmes-verdict-theranos-founder-is-guilty-on-four-of-11-charges-in-fraud-trial-11641255705?mod=hp_lead_pos1. Accessed: 2022-01-04. 2022.

Scharff, R.C. *Heidegger Becoming Phenomenological: Interpreting Husserl through Dilthey, 1916–1925*. New Heidegger Research. Rowman & Littlefield International, 2018. ISBN: 9781786607744.

— *How History Matters to Philosophy: Reconsidering Philosophy's Past After Positivism*. Routledge Studies in Contemporary Philosophy. Taylor & Francis, 2014. ISBN: 9781134626731.

Scott, Jason. *A Lancaster Teen Remembers BBSes (February 6, 2008)*. http://textfiles.com/history/alancasterteen.txt. Accessed: 2021-12-14. 2008.

— *BBS Documentary Interview: Jayne*. https://archive.org/details/bbs-20030523-jayne. Accessed: 2021-11-08. 2003.

— *Phone Phreaking*. http://textfiles.com/phreak/. Accessed: 2021-12-14. 2021.

— *The Text of my 1999 DEFCON 7 Speech*. http://www.textfiles.com/thoughts/speech.txt. Accessed: 2021-11-08. 1999.

Sejnowski, T.J. *The Deep Learning Revolution*. The MIT Press. MIT Press, 2018. ISBN: 9780262038034.

Serrano, Sofia and Noah A Smith. "Is Attention Interpretable?" In: *Proceedings of ACL*. 2019.

Sharma, Shubham, Jette Henderson, and Joydeep Ghosh. "CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 166–172. ISBN: 9781450371100. DOI: 10.1145/3375627.3375812. URL: https://doi.org/10.1145/3375627.3375812.

Shefsky, Jay. *30 Years Later, Notorious 'Max Headroom Incident' Remains a Mystery*. https://news.wttw.com/2017/11/21/30-years-later-notorious-max-headroom-incident-remains-mystery. Accessed: 2021-11-22. 2017.

Silicon Valley Historical Association. *Steve Jobs Interview about the Blue Box Story*. https://www.youtube.com/watch?v=HFURM8O-oYI. Accessed: 2021-11-17. 1994.

Smithsonian National Museum of American History. *Altair 8800 Microcomputer*. https://americanhistory.si.edu/collections/search/object/nmah_334396. Accessed: 2021-12-13. 2021.

Stallman, Richard. *On Hacking*. http://www.stallman.org/articles/on-hacking.html. Accessed: 2021-11-08. 2000.

Sterling, Bruce. *The Hacker Crackdown*. https://www.mit.edu/hacker/hacker.html. Accessed: 2021-11-08. 1994.

Stone, B. *The Everything Store: Jeff Bezos and the Age of Amazon*. Little, Brown, 2013. ISBN: 9780316219259.

The Museum of Classic Chicago Television. *WGN Channel 9 - The first Max Headroom Incident*. https://www.youtube.com/watch?v=dKnwhokvgxE. Accessed: 2021-11-22. 1987.

— *WTTW Chicago - The Max Headroom Pirating Incident*. https://www.youtube.com/watch?v=cycVTXtm0U0. Accessed: 2021-11-22. 1987.

The Top 500 Project. *Operating System Share*. https://www.top500.org/. Accessed: 2021-12-24. 2021.

Tiku, Nitasha. *Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it.* https://www.washingtonpost.com/technology/2020/12/23/google-timnit-gebru-ai-ethics/. Accessed: 2022-03-18. 2020.

Torvalds, Linus. *Linux's History*. https://www.cs.cmu.edu/~awb/linux.history.htmlt. Accessed: 2021-12-14. 1992.

Tozzi, C. and J. Zittrain. *For Fun and Profit: A History of the Free and Open Source Software Revolution*. History of Computing. MIT Press, 2017. ISBN: 9780262341189.

Van Esch, Patrick, J Stewart Black, and Joseph Ferolie. "Marketing AI recruitment: The next phase in job application and selection". In: *Computers in Human Behavior* 90 (2019), pp. 215–222.

Vaswani, Ashish et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

Villa-Nicholas, M. *Latinas on the Line: Invisible Information Workers in Telecommunications*. Latinidad: Transnational Cultures in the United States. Rutgers University Press, 2022. ISBN: 9781978813731.

Vincent, James. *What a machine learning tool that turns Obama white can (and can't) tell us about AI bias*. https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias. Accessed: 2022-04-13. 2020.

Vollmann, Michael T. *The 414s: The Original Teenage Hackers*. https://vimeo.com/502242358. Accessed: 2022-01-14. 2015.

Wachter, Sandra, Brent D. Mittelstadt, and Chris Russell. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR". In: *CoRR* abs/1711.00399 (2017). arXiv: 1711.00399. URL: http://arxiv.org/abs/1711.00399.

Wang, Yilun and Michal Kosinski. "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images." In: *Journal of personality and social psychology* 114.2 (2018), p. 246.

Ward, Marguerite. *Mark Zuckerberg returns to the Harvard dorm room where Facebook was born*. https://www.cnbc.com/2017/05/25/mark-zuckerberg-returns-to-the-harvard-dorm-where-facebook-was-born.html. Accessed: 2022-02-19. 2017.

Weber, M. *The Protestant Ethic and the Spirit of Capitalism*. Routledge Classics. Taylor & Francis, 2005. ISBN: 9781134521883.

Winner, Langdon. "Do Artifacts Have Politics?" In: *Daedalus* 109.1 (1980), pp. 121–136. ISSN: 00115266. URL: http://www.jstor.org/stable/20024652.

Woodward, James. *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in Philosophy of Science. Oxford University Press, 2003. ISBN: 9780198035336.

World Wide Web Technology Surveys. *Usage statistics of operating systems for websites*. https://w3techs.com/technologies/overview/operating_system. Accessed: 2021-12-24. 2021.

Wozniak, Steve. *Steve Wozniak on What Really Happened in Jobs' Garage*. https://www.bloomberg.com/news/videos/2014-12-05/steve-wozniak-on-what-really-happened-in-jobs-garage. Accessed: 2022-02-19. 2014.

Wu, Xiaolin and Xi Zhang. *Responses to Critiques on Machine Learning of Criminality Perceptions (Addendum of arXiv:1611.04135)*. 2017. arXiv: 1611.04135 [cs.CV].

Yang, Diyi et al. "Who Did What: Editor Role Identification in Wikipedia." In: *Proceedings of the AAAI Conference on Web and Social Media (ICWSM)*. 2016, pp. 446–455.

Youth International Party. *SchoolStoppers Textbook*. http://textfiles.com/100/killshco.ana. Accessed: 2022-02-07. 1983.

Zaw Thiha Tun. *Theranos: A Fallen Unicorn*. https://www.investopedia.com/articles/investing/020116/theranos-fallen-unicorn.asp. Accessed: 2022-03-04. 2022.

Zwick, Rebecca. "Is the SAT a 'Wealth Test'?" In: *Phi Delta Kappan* 84.4 (2002), pp. 307–311. DOI: 10.1177/003172170208400411. eprint: https://doi.org/10.1177/003172170208400411. URL: https://doi.org/10.1177/003172170208400411.

Vita

Christopher Grimsley

Place of Birth:

- Baltimore, MD

Education:

- University of Kentucky, Lexington, KY
  M.A. in Philosophy, Dec. 2019
- Frostburg State University, Frostburg, MD
  M.A. in Teaching, May. 2010
- Frostburg State University, Frostburg, MD
  B.S. in Sociology, May. 2007

Professional Positions:

- Graduate Teaching Assistant, University of Kentucky Fall 2017–Spring 2022

Honors

- Outstanding Teaching Assistant Award, University of Kentucky College of Arts & Sciences. April, 2020.

Publications & Preprints:

- Grimsley, Christopher, Elijah Mayfield, and Julia R.S. Bursten. "Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models." In Proceedings of The 12th Language Resources and Evaluation Conference, 1780–1790. Marseille, France: European Language Resources Association, 2020. `https://www.aclweb.org/anthology/2020.lrec-1.220`.