Wittgenstein, Natural Language Processing, and Ethics of Technology

Undergraduate Honors Thesis

Submitted to Department of Philosophy

Duke University, April 12, 2020

Table of Contents

Introduction

Natural Language Processing (NLP) is a field of research and set of technologies that aims to improve how computers interact with human language. Some common applications of NLP include converting audio into text to execute spoken commands (speech-to-text and voice assistants); retrieving information from a text (text and sentiment analysis); and generating natural language in conversation (chatbots and predictive text). Google Translate, Apple's Siri, Amazon's Alexa, Microsoft Word Editing, and Grammarly are all technologies which use NLP [1]. As a subset of Artificial Intelligence (AI), NLP uses a wide variety of methods, from statistical models to machine learning to deep learning neural networks. As I explain in Chapter One, these methods convert words into numbers, use computational models to do math with the numbers, and apply them in a variety of technologies.

NLP is an illuminating case study for anyone interested in AI, technology, and language. In this thesis, I show that using Wittgenstein's work to investigate NLP has distinct advantages. First of all, Wittgenstein takes seriously the fact that most mathematicians and computer scientists who work on NLP could care less about philosophy[2]. NLP technologies and models are developed for specific, practical applications and are not intended to be models of language as such. It might be tempting to draw out which "theories of language" have been "proven" in NLP, but what counts as a successful speech-to-text or predictive text software will not necessarily count as a successful theory of language. Thus, I don't see NLP as an applied philosophy of

[1] Robert, Dale. "The commercial NLP landscape in 2017." *Natural Language Engineering*. Cambridge University Press, Vol. 23, Issue. 4, July 2017, pp. 641-647. Accessed from https://www.cambridge.org/core/journals/natural-language-engineering/article/commercial-nlp-landscape-in-2017/CDCE7C93C48CFDF9094908F3A0DB9E26

[2] Wilks, Yorick. "A Wittgensteinian computational linguistics?" *Proceedings of the AISB 2008 Symposium on Computing and Philosophy*, 2008, p. 2.

language, but as an example which can illuminate how technology is intertwined with language and philosophy.

The second reason I draw on Wittgenstein is that, perhaps ironically, his ideas are being credited for inspiring major breakthroughs in NLP and AI research. As I describe in Chapter Two, a growing number of people compare contemporary NLP methods like machine learning and Word2Vec to Wittgenstein's ideas of "meaning-as-use," "family resemblances," and "blurry concepts." I should emphasize again that the significance of these comparisons isn't to put forth arguments for a Wittgensteinian "theory" of language as such. This would be antithetical to Wittgenstein's entire spirit of inquiry and description. My concern is not whether NLP methods show how Wittgenstein was right or wrong about language, but what to do with the fact that a certain interpretation of Wittgenstein is gaining traction in NLP. At the very least, these comparisons allow me to clarify some misconceptions about Wittgenstein, as I do at the end of Chapter Two. Furthermore, the comparison between Wittgenstein and NLP opens the door for me to argue that a Wittgensteinian *spirit* - one which re-centers the human and social aspects of technological development - is needed in NLP and AI. In Chapter Three, I use Wittgenstein's work to gain clarity on the ethics of technology and AI. I show that we are "held captive" by a "picture" of technology, to use Wittgenstein's words (§115)[3] , in which the technical and the human/social/ethical belong to two completely separate worlds. To demonstrate this, I look closer at the way we talk and write about technology and AI.

Throughout this thesis, I draw on a broad network of Wittgensteinian inspired work by Juliet Floyd, Toril Moi, and Stanley Cavell. I put this so-called "Ordinary Language Philosophy" in conversation with scholars at the intersection of science and race, including Denise Ferreira da

[3] Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe, P. M. S. Hacker, Joachim Schulte. Wiley-Blackwell. 2009, §115.

Silva, Ruha Benjamin, Stephanie Spoto, Frank Wilderson, and Louis Chude-Sokei. My primary goals in this thesis are to give a "surveyable representation" (§122)[4] of the landscape of NLP, the ethics of technology and AI, and to put Wittgenstein in conversation with scholars of race and technology.

Chapter One

In this chapter, I give an overview of the history of NLP and contemporary NLP methods. I discuss the transition from "rule-based" methods which use traditional "if-then" algorithms, to statistical methods which apply probabilistic models to data, and finally to "data-driven" methods such as machine learning and neural networks which make their own rules using data. . I describe NLP methods such as word-embeddings, Latent Semantic Analysis (LSA), and Word2Vec. I then turn my attention to how these methods and technologies are situated in practices of surveillance, labor exploitation, and misleading marketing

History of NLP

Like many technologies, NLP has roots in the military-industrial complex. Around the end of World War II, the United States military paid researchers at Georgetown University, MIT, and IBM to develop a "machine translation" tool which could translate Russian and German intelligence documents into English[5]. At first glance, machine translation seemed simple enough. First, researchers would rearrange Russian and English sentences into a format a machine could use - standardizing a sentence's grammatical structure, coding grammar conversion rules[6], and

[4] Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe, P. M. S. Hacker, Joachim Schulte. Wiley-Blackwell. 2009, §122.
[5] Reifler, Erwin. "Machine translation." *Bulletin of the Medical Library Association,* vol. 50, no 3, 1962, pp. 473-80.
[6] Reifler, Erwin. "Machine translation." 1962, pp. 473-80.

equipping a machine with a Russian-to-English dictionary. This rule-based approach to language was influenced by the likes of Frege, Russell, Carnap[7], and linguist and philosopher Noam Chomsky[8]. In his 1957 book "Syntactic Structures," Chomsky argued that sentences could be organized in "Phase-Structure Grammar" and computers could follow "if-then" statements to determine their meaning[9]. For example: *If* the sentence begins with "Heil Hitler," *then* it is a German Nazi intelligence report. This method had some success in decoding encrypted messages, encouraging the belief that human language was yet another complicated code to be cracked[10]. As a result, machine translation seemed even more readily achievable. If the human language was a system of symbols governed by some set of rules, researchers were confident that a machine could convert Russian to English at a faster and cheaper rate than human translators. As one NLP researcher stated in 1952, "there is a simple mechanical solution for the...complete elimination of the human in machine translation." [11]

By the beginning of the Cold War, however, it became clear that this rule-governed approach to machine translation was not proving successful. As one NLP researcher put it in 1962, the complexity of language presented "paramount linguistic problems" for the use of "electronic calculators for the mechanization of the translation process."[12] NLP researchers had to essentially brain-storm into existence every rule about syntax, grammar, and language use. Finding ways to tame the complex "symbolization of language" and "the problem of multiple

[7] Wilks (2008) argues that Carnap's 1936 paper "The Logical Syntax of Language" is the unacknowledged but nevertheless essential groundwork for Chomsky's work on universal grammar and transformation rules.)Rudolf Carnap was Noam Chomsky's teacher.) In the 1936 paper, Carnap puts forth a system which can separate meaningful from meaningless expressions by means of rules. For more information, see: Carnap, Rudolf. "Logische Syntax der Sprache," 1936. Translated by Kegan Paul. London, 1937.
[8] Wilks, Yorick. "A Wittgensteinian computational linguistics?" 2008, p. 2.
[9] Chomsky, Noam. *Syntactic Structures*, Mouton, The Hague, 1957.
[10] Du Pont, Quinn. "The Cryptological Origins of Machine Translation." *Amodern 8*, 2018. Accessed from https://amodern.net/article/cryptological-origins-machine-translation/
[11] Reifler, Erwin. "Machine translation." 1962, p. 479.
[12] Reifler, Erwin. "Machine translation." 1962, p. 474

meanings"[13] stumped NLP researchers. The additional technical challenge of making these sentences intelligible for low-power processing machines was also daunting. The 1960's was still the era of punch cards; in comparison to today's computers, these "electronic calculators" had virtually no power, no memory, and no advanced programming language[14]. Even after advancements in programming language and answer-question systems, NLP was still not a feasible tool on a large scale. In 1966, the Automatic Language Processing Advisory Council (ALPAC) released a pessimistic report that effectively ended twelve years of research into machine translation[15]. In the report's own words:

> Computers give promise of helping us control the problems relating to the tremendous volume of data, and to a lesser extent the problems of data complexity. But we do not yet have good, easily used, commonly known methods for having computers deal with language data[16].

Accompanied by a recommendation to decrease government funding, the ALPAC report effectively marked the beginning of the "Ice Age" of NLP[17].

In the 1970's, the field of natural language processing was reinvigorated by improved computer capabilities and the commercialization of information. While the military had driven NLP to focus on machine translation, private companies used NLP to navigate industry-specific databases, automate administrative tasks, and serve various business functions[18]. Accordingly, the goals of NLP shifted from complete comprehension of an entire language to partial comprehension for the sake of a specific task, such as information retrieval or the automatic

[13] Reifler, Erwin. "Machine translation." 1962, p. 476.
[14] "Timeline of Computer History." *Computer History Museum.* Accessed from https://www.computerhistory.org/timeline/1960/
[15] "Automatic Language Processing Advisory Committee. "Language and Machines. Computers in Linguistics and Translation." 1966, p. 14. Accessed from http://www.mt-archive.info/ALPAC-1966.pdf.
[16] Automatic Language Processing Advisory Committee. "Language and Machines. Computers in Linguistics and Translation." 1966, p. 14. Accessed from http://www.mt-archive.info/ALPAC-1966.pdf.
[17] Lanzetta, Michael. "Machine Learning, Deep Learning, and Artificial Intelligence." *Artificial Intelligence of Autonomous Networks.* Edited by Mazin Gilbert. CRC Press, 2018, section 2.2.1.
[18] Robert, Dale. "The commercial NLP landscape in 2017." 2017.

digitization of text. In turn, NLP was especially popular in information-driven industries such as advertising and marketing, foreshadowing the success of Google and Facebook as today's biggest advertising companies. For example, in the 1970's, secretaries at the marketing firm J. Walter Thompson developed some of the first NLP technologies dedicated to consumer intelligence[19]. They transcribed and digitized newspaper articles, consumer reports, and customer reviews; classified and organized this information into databases; and wrote functional code to search for keywords, track market trends, and automate administrative functions[20]. Because these NLP databases didn't need to be "scaled up" to process the entire English language - as with machine translation - these "expert systems" could proceed with rule-based models which were hand-coded by experts in the field.

In the 1980's and 1990's, researchers decided to revisit the challenge of general language comprehension, returning to statistical methods used by Andrey Markov and Alan Turing in the 1970's[21]. This time around, however, these approaches were technically supported by rapidly increasing computational power and enlarged samples of data. As a result, the last of Chomskyian rule-based methods began to give way to the data-driven empiricism of statistics and machine learning[22]. In the late 1990's, the rise of the World Wide Web ushered in a large, shared corpus (or collection) of language-data[23]. Since then, the era of "Big Data" has spurred rapid developments in NLP. Using invasive surveillance techniques and the "micro-labor" of

[19] Luksic, Sandra. "Technology, Advertising, and Women." *Consuming Women, Liberating Women: Women and Advertising in the Mid-Twentieth Century*. 2019. Accessed from https://sites.duke.edu/womenandadvertising/exhibits/tech-ads-and-women/
[20] Luksic, Sandra. "Technology, Advertising, and Women." 2019.
[21] Monica Franzese, Antonella Iuliano. "Hidden Markov Models." *Encyclopedia of Bioinformatics and Computational Biology*, Volume 1, 2019, Pages 753-762. Accessed from https://www.sciencedirect.com/science/article/pii/B9780128096338204883
[22] Sparck Jones, Karen. "Natural Language Processing: a historical review." *Computer Laboratory, University of Cambridge*, 2001. Accessed from https://www.cl.cam.ac.uk/archive/ksj21/histdw4.pdf
[23] Sparck Jones, Karen. "Natural Language Processing: a historical review." 2001.

thousands of outsourced workers, large samples of "language data" have been collected and classified[24]. The human language has been categorized in a variety of ways, from being organized by concept to being compiled in "dictionaries" with their definitions, parts of speech, syntax rules, and more. The availability and sheer amount of online text has been essential for building the contemporary NLP technologies we use today.

Contemporary NLP

While many contemporary NLP methods make some use of rule-based approaches - especially in industry specific datasets - these rules are almost always situated within a more flexible model driven by probabilities and statistics. One basic statistical approach that was popularized in the 1990's is the count-based approach[25]. Two very simple versions are bag-of-words models and n-gram models. As the name suggests, bag-of-words models treat text like a bag of words: they don't care about the order of words or their context, the model just looks at how frequently words occur. However, because the same words can be rearranged to produce different meanings, bag-of-words models are less frequently used than n-gram models. N-gram models don't divide sentences into individual words, but into pairs (bi-grams), trios (tri-grams), and so forth. For example, the sentence "The quick brown fox jumps over the lazy dog" can be divided into the trigrams "the quick brown," "quick brown fox," "brown fox jumps," and so forth. A common use of n-gram models is to determine the likelihood of something given a certain sequence. For example, if a sentence contains two of the above trigrams, then the model might correctly predict the rest of the sentence to be "The quick brown fox jumps over the lazy dog."

[24] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 5. Accessed from https://arxiv.org/pdf/2001.05046.pdf

[25] Laponin, Arne. "The theory you need to know before you start an NLP project." *Towards Data Science.* 2019. Accessed from https://towardsdatascience.com/the-theory-you-need-to-know-before-you-start-an-nlp-project-1890f5bbb793

To use these methods, a body of text has to be "cleaned" and pre-processed[26]. One common first step is breaking up sentences into individual words or n-grams, which is called "tokenizing."[27] For example, "I want Bernie Sanders to become president," becomes a list of words: I, want, Bernie, Sanders, to, become, president. In addition to removing punctuation and spaces, a data scientist might also choose to remove "stop words," such as *the, and, your, a,* and *it.* We get: want, Bernie, Sanders, become, president. Often, a computer scientist will also "stem" words, reducing different versions of the same word to one word, like changing *winning, winners, won, winnings* to *win*[28]. Likewise, computer scientists will spell check words and standardize synonyms, for instance, choosing either "Palestine" or "Israeli-occupied land." Using a combination of humans and automated algorithms, words are classified, defined, and organized. Researchers may label each word with a part of speech, definition, sentiment, or other characteristic they want to know about it. For example, in one sentiment analysis dictionary used to track the emotions of words, "Black" is labeled "negative, fear, and sadness" while "White" is labeled "positive, anticipation, joy."[29] Another database called "WordNet" organizes words into nested categories. For example, "chair" is classified as Artifact > furnishing > furniture > seat[30]. Both datasets were made using Amazon's "Mechanical Turk," a crowdsourcing website in which researchers pay people to complete small tasks such as labeling a word or sentence with a definition or emotion[31].

[26] Koenig, Rachel. "NLP for Beginners: Cleaning and preprocessing text data." *Towards Data Science,* 2019. Accessed from https://towardsdatascience.com/nlp-for-beginners-cleaning-preprocessing-text-data-ae8e306bef0f
[27] Laponin, Arne. "The theory you need to know before you start an NLP project." 2019.
[28] Laponin, Arne. "The theory you need to know before you start an NLP project." 2019.
[29] These labels can be found in the automated NRC Hashtag Emotion Lexicon. Accessed from https://saifmohammad.com/WebPages/AccessResource.htm
[30] Crawford, Kate and Paglen, Trevor. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019. Accessed from https://www.excavating.ai/
[31] "About Amazon Mechanical Turk." *Amazon Mechanical Turk: Frequently Asked Questions.* Accessed from https://www.mturk.com/worker/help

Companies like Amazon, Apple, Facebook and Google have been essential for the development of NLP, and in particular for machine learning methods. In order to understand these methods, it's necessary to give an overview of what Shoshanna Zuboff calls "surveillance capitalism," or the commodification of personal information for profit[32]. Since the development of the internet, *billions* of conversations, comments, arguments, articles, Tweets, videos, recordings, and images have been downloaded or "scraped" off of the internet[33]. Without adequate knowledge or consent[34], your "language data" gets used to fuel NLP methods. These methods are then applied in technologies such as Amazon Alexa, Google Home, iPhone, and Amazon, which serve as these companies' eyes and ears.

Without your personal data, it is likely that the contemporary state of modern machine learning and deep learning would not be possible; Google and Facebook would simply not have enough material to train and test their models on[35]. Similarly, the human language would not be structured in ways that models could use without massive amounts of human labor and labor exploitation. For example, the task of pre-classifying data to train supervised machine learning models falls on exploited laborers in online shadow economies[36]. As mentioned, corporations like Amazon Mechanical Turk pay people mere pennies to perform small tasks such as tagging words and sentences with definitions, emotions, and parts of speech. But this labor happens outside of official working conditions too. For example, if you have ever been asked to complete a re-CAPTCHA or to "Select all the images which contain a vehicle," then you have done unpaid and coerced labor for Google's hand-writing to text software and their self-driving car projects,

[32] Zuboff, Shoshana. "Surveillance Capitalism." *Project Syndicate.* Accessed from https://www.project-syndicate.org/onpoint/surveillance-capitalism-exploiting-behavioral-data-by-shoshana-zuboff-2020-01?barrier=accesspaylog

[33] "Scraped" is a fancy word for capitalism eavesdropping on everything you say.

[34] Wilks, Yorick. "A Wittgensteinian computational linguistics?" 2008, p. 2.

[35] Birhane and Van Dijk. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 5.

[36] Birhane and Van Dijk. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 5.

all under the guise of proving "I am not a robot" to a company who is pretending to replace you with one[37]. Similarly, Facebook checks its voice-to-text software by hiring people to transcribe audio taken from your video conversations on Facebook Messenger[38]. Although Facebook comes under fire for allowing harmful content on its platform, they still hire thousands of workers to remove the worst of the worst posts and videos - a draining job which involves reading thousands of vicious slurs, death threats, and rape fantasies for minimum wage and no health care benefits[39]. Even on the better-paid side of things, "cleaning" or pre-processing language data, such as removing stop words and stemming words, requires data scientists to carefully consider the data and its potential use cases. This is all to highlight that the same technical methods which advance NLP methods can also perpetuate social and economic inequality. The more "automated" a technology is, the more human labor is simultaneously required and erased[40].

In professional settings and computer science classes[41], however, machine learning refers to a model or algorithm which "automatically" updates its rules for making decisions[42]. The basic goal of machine learning is to teach an algorithm to classify or predict data it has never seen. There are two ways to "train" a machine learning model: supervised and unsupervised learning. In supervised learning, computer scientists show the computer pre-labeled data. In

[37] Schmieg and Lorusso. "Five Years of Captured Captchas." 2017. Accessed from http://five.yearsofcapturedcapt.ch/as

[38] Frier, Sarah. "Facebook Has Been Paying Contractors to Transcribe Users' Facebook Messenger Voice Chats." *Time Magazine.* 2019. Accessed from https://time.com/5651395/facebook-contractors-transcribe-conversations-audio-files/?xid=tcoshare#0e3bc275-fd0c-4675-9a34-1d012af032c3?utm_source=twitter.com&utm_medium=social&utm_campaign=social-share-article

[39] Wong, Queenie. "Facebook content moderation is an ugly business. Here's who does it." *CNET.* 2019. Accessed from https://www.cnet.com/news/facebook-content-moderation-is-an-ugly-business-heres-who-does-it/

[40] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 5. Accessed March 2020 from https://arxiv.org/pdf/2001.05046.pdf

[41] Fiesler, Casey. "What Our Tech Ethics Crisis Says About the State of Computer Science Education." *Next.* December 5, 2018. Accessed from https://howwegettonext.com/what-our-tech-ethics-crisis-says-about-the-state-of-computer-science-education-a6a5544e1da6

[42] "What is Machine Learning?" *IBM.* Accessed from https://www.ibm.com/topics/machine-learning

unsupervised learning, the data has fewer labels. After feeding the machine this training data, the algorithm is tested on new data[43]. For example, say we wanted to make a supervised training model to identify which incoming emails in my inbox are spam. First, we might hire workers on Amazon Mechanical Turk to classify thousands of emails as "spam" or "not spam." Then, we "train" the model by showing it this pre-classified "training" data. Once the model learns how to predict the difference between regular and spam emails, it is tested on a new set of *unclassified* "testing" data.

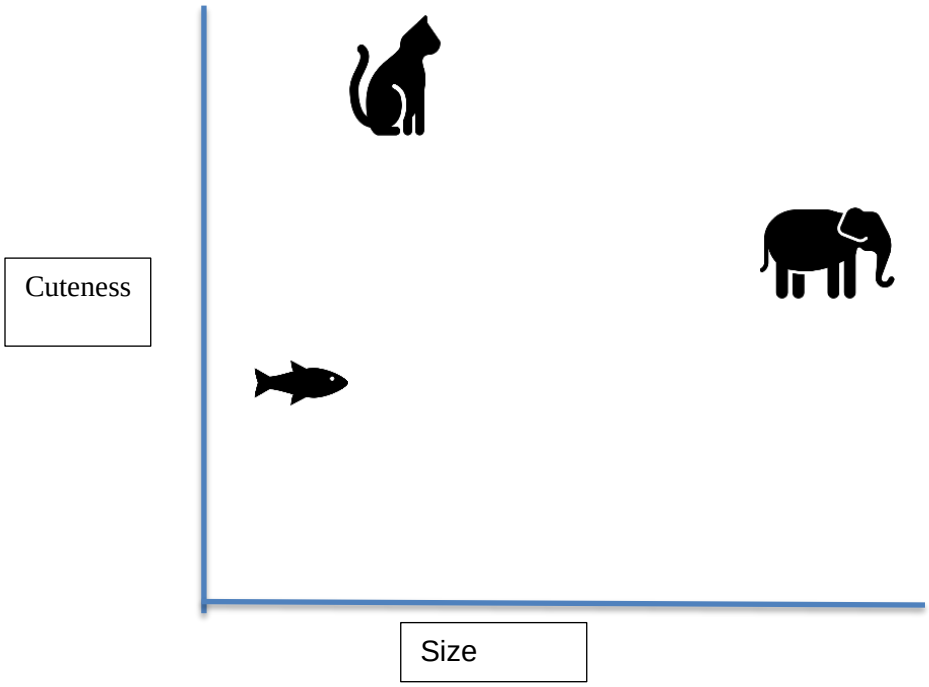
How do machine learning algorithms "learn" how to classify spam emails from regular ones? This process is called the "weighting process," in which the algorithm picks out the most predictive features in the data. Think of the weighting process as being part of a jury in a courtroom trial. The jury is told to make a guilty or not guilty decision according to some set of criteria laid out by the judge and lawyers. The jury must consider which evidence and testimony is most important, theoretically changing their minds as the new evidence is presented. In a more technical sense, the algorithm uses complex mathematical modeling to see which features in the data should be given the most weight. Then, it uses those weights to produce a prediction about some feature. If the prediction is off, the algorithm may or may not adjust the weights.

Understanding human language is complex enough as it is, so how do machine learning methods help computers predict the meaning of a word? To answer this question, let's walk through an example that uses two advanced machine methods called Latent Semantic Analysis

[43] There are two common tasks a machine learning algorithm is tested on: regression and classification. The answer to classification tasks is either a yes or a no - something belongs to the category or it doesn't. A regression model, on the other hand, might tell us the probability of something belonging to a category.

(LSA) and Word2Vec[44]. LSA uses something called *word-embeddings*[45]. The basic idea behind word-embeddings is that any document, phrase, n-gram, word, or letter can be turned into a number - and once we turn words into numbers, we can do math. For instance, we could turn the words for cat, elephant, and goldfish into word-embedding by ranking them on a scale from 1-10 in size and cuteness. A cat might be represented by the number (3, 10) an elephant (9.5, 6,) and a fish (1, 7) Then, we can chart and graph these word-embeddings like this:

Animal	Size	Cuteness
Cat	3	10
Elephant	9.5	6
Fish	1	3



[44] This example is based on a 2017 research paper titled "Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database." In this study, researchers compared two methods for "topic modeling", or uncovering the "structure" of a text's data in order to gain information about that text's topics. For more information, see: Altszyler, Edgar et al. "The Interpretation of Dream Meaning: Resolving Ambiguity Using Latent Semantic Analysis in a Small Corpus of Text." *Consciousness and Cognition 56*, 2017, pp. 178–187.
[45] Brownlee, Jason. "What are Word Embeddings?" *Machine Learning Mastery*. 2017. https://machinelearningmastery.com/what-are-word-embeddings/

Looking at this graph, we might conclude that a fish and cat are similar in size, but not in cuteness. Likewise, we could use the graph to make predictions about how cuteness and size relate, for instance, predicting that the bigger an animal is, the less cute it is. When looking at this graph, an intuitive assumption about word-embeddings is revealed: points which are close to each other are also supposed to be similar. This assumption is known as the Distributional Hypothesis[46], and it's why we can look at word embeddings and infer meaning the way we do.

Word embeddings can be understood as vectors, or a list of numbers which represent certain features about some text. In LSA, the process of creating vectors sometimes involves creating a giant spreadsheet called a "Document Term Matrix." This matrix keeps track of which words appear in which document. For example, if a word is used in a document, we might indicate it with 1, but if it doesn't, it gets a 0. The vectors for each document are then represented by a list of zeroes and ones; for instance, (0, 0, 0, 0, 1). Now, we can do math with these vectors. For instance, if we think of these vectors as coordinates, we can map them in a high-dimensional space. Drawing on the Distributional Hypothesis, we can compute the similarity between two documents by calculating how close their two vectors are. Likewise, we can also use machine learning methods to predict the meaning of an unknown document by treating these vectors as training and testing data. Recall that unsupervised machine learning models looks at the underlying structure of data rather than referring to a pregiven classifications – this "structure" could be the mathematical relations between vectors. Then, given this data, the goal of LSA would be to find the underlying patterns or clusters in different documents, thereby revealing "latent" topics or clusters in a large group of text [47].

[46] Altszyler, Edgar et al. "The Interpretation of Dream Meaning: Resolving Ambiguity Using Latent Semantic Analysis in a Small Corpus of Text." *Consciousness and Cognition 56*, 2017, pp. 178–187.
[47] Altszyler, Edgar et al. "The Interpretation of Dream Meaning: Resolving Ambiguity Using Latent Semantic Analysis in a Small Corpus of Text." 2017, pp. 178–187.

Another contemporary NLP method is Word2Vec. Word2Vec is often used to graph individual words or n-grams, not entire documents.A basic example which shows the logic of Word2Vec is King - Man + Woman = Queen. But Word2Vec also turns words into numbers so we get to unlock the power of math. For instance: King(10, 5, 7) - Man(5, 8, 1) + Women(1, 3, -9) = (6, 0, -3) in which (6, 0, -3) represents a vector that is close to the words for Queen. Unlike LSA, these numbers do not come from a Document Term Matrix, but are created using advanced machine learning methods[48]. Before I explain more about Word2Vec, it's important to note the actual numbers in vectors don't matter so much as the way they relate to each other. In theory, there are endless ways to turn text into numbers. For example, we could list words alphabetically, from A to Z, and have each vector represent the order of each word. The word "a" might be (1,0,0,0,0,0) while "police" would be (0,0,0,0,1,0). Another way to create vectors is to create a matrix that includes more information about the word. Each number in this matrix could say something about how a word is used across multiple documents, such as how frequent it is, what other words it appears with, and so forth. When we use these methods, Word2Vec vectors are sometimes called "distributed representations" because they tell us how one word is distributed throughout a text[49].

In advanced cases, Word2Vec uses "deep learning" methods called "neural networks" to model the relationships between words[50]. Neural networks are very loosely inspired by the neurons in the brain. The key idea is that when some neurons fire, they activate other neurons, which activate other neurons, and the neuron which fires the strongest at the end is the chosen

[48] Altszyler, Edgar et al. "The Interpretation of Dream Meaning: Resolving Ambiguity Using Latent Semantic Analysis in a Small Corpus of Text." 2017, pp. 178–187.
[49] Altszyler, Edgar et al. "The Interpretation of Dream Meaning: Resolving Ambiguity Using Latent Semantic Analysis in a Small Corpus of Text." 2017, pp. 178–187.
[50] "A Beginner's Guide to Word2Vec and Neural Network Embeddings." *Pathmind.* Accessed from https://pathmind.com/wiki/word2vec

output. Neural networks are made of an input layer, a few "hidden" layers, and an output layer. The basic goal is to put our raw text into the input layer and get some kind of prediction from the output layer. As this process is repeated, the "neurons" get better and better at figuring out just *how* they should activate each other to produce the most accurate output, finding the best rules for decision making[51]. Once the neural network has been adequately trained, the hidden layer produces a mathematical model known as a "learned linear transformation," which becomes the most likely numerical representation for that word[52].

In Word2Vec, we use two types of neural networks called Continuous Bag of Words (CBOW) and Skip-Gram[53]. The goal of CBOW is to fill in a blank i.e. to predict a word vector from its context vectors. For example, if we had the sentence "running from police," the CBOW model uses "running" and "police" to predict the middle word "from." On the other hand, the goal of Skip-Gram is to predict acceptable contexts for a word. A Skip-Gram model is trained to predict "running" and "police" from the word "from." CBOW and Skip-Grams are used to train a neural-network which can both predict the best output vector and find the correct - or rather, the most likely - neighboring vectors. The purpose of using Skip-Gram in conjunction with CBOW is to train our neural network model to predict the meaning of both old words in new contexts and place new words in their right contexts.

Of course, there are countless other ways to build NLP that I cannot explain in this thesis. But before I describe how the methods I *have* covered relate to Wittgenstein's work, I want to

[51] This adjustment process involves finding the difference between the predicted vector and the actual vector. This difference is used to calculate the "error" vector. The "error vector" is helpful information for the model and used to adjust the activation process in the inner layers. In other words, the model learns from its mistakes to create a better path moving forward.

[52] "A Beginner's Guide to Word2Vec and Neural Network Embeddings." *Pathmind.* Accessed from https://pathmind.com/wiki/word2vec

[53] "A Beginner's Guide to Word2Vec and Neural Network Embeddings." *Pathmind.* Accessed from https://pathmind.com/wiki/word2vec

recap what I have said so far. With Big Data NLP methods, and especially unsupervised machine learning, the algorithm "learns" the meaning of a word not by looking at its place in some classification schema, but by predicting its use in a data set. For example, word embeddings turn words into numbers based on how similar they are to other words, and Word2Vec creates vectors based on a word's context. In NLP, the similarity between two words is often thought of as a statistical relationship, such as how often they co-occur with each other. According to the Distributional Hypothesis, the closer two vectors are, the more similar the meaning of those words. While Word2Vec doesn't exactly follow the Distributional Hypothesis, it predicts a word's meaning by looking at its neighboring data points. As I will explore next, all of these methods, in one way or another, have been compared to Wittgenstein.

Chapter 2

Introduction

In the past few decades, dozens of scholarly articles, computer science forums, and GitHub tutorials have drawn attention to Wittgenstein's relevance to NLP. The nature of this small but growing interest can be characterized by Graeme Hirst's semi-ironic statement that "The solution to any problem in AI may be found in the writings of Wittgenstein, though the details of the implementation are sometimes rather sketchy."[54] Initially, I was thrilled to discover a small but growing community of people who also saw the similarities between contemporary NLP methods and Wittgenstein's work. From a technical perspective, the similarities are striking. For instance, Wittgenstein's discussion of "family resemblances" can be aptly applied to

[54] Wilks, Yorick. "Wittgenstein and Computational Linguistics." 2008, pg. 5. Accessed from https://www.academia.edu/8452891/Wittgenstein_and_Computational_Linguistics

unsupervised learning classification techniques such as cluster analysis. Others have argued Wittgenstein's idea of "meaning as use" can be found in Word2Vec, which learns the meaning of a word through anything *but* its definition. In many ways, it is remarkable how the *Tractatus* and *Philosophical Investigations*, published in 1945 and 1953, can inform the challenges which NLP faces today.

However, the more I reflect on these comparisons, the more they disappoint me. Wittgenstein focuses on how language is an activity between humans. NLP, as a profit-driven technology, tries to make machines process language while suppressing and disguising the human activity and judgment involved in this processing. What makes Wittgenstein so useful for the technical development of NLP is also what puts his work in stark contrast with the goals of corporations who weaponize Big Data. By taking Wittgenstein's ideas from a purely technical perspective, NLP has flattened his emphasis on particularity and realism to build computational models driven by Big Data and empiricism.

In this chapter, I will give an overview of the uptake of Wittgenstein's work by practitioners of NLP. I will focus on Wittgenstein's relationship to three components of contemporary NLP: Big Data, Word2Vec, and cluster analysis. Along the way, I will expand on Wittgenstein's ideas and point out further connections I have made for myself. Then, I will show why these technical interpretations - while often striking - fall short of Wittgenstein's revolutionary potential. Finally, I will touch on how what these misreadings can say about our picture of technology.

NLP and Wittgenstein: Similarities

In "A Wittgensteinian computational linguistics?[55]," and "Wittgenstein and Computational Linguistics[56]," NLP researcher Yorick Wilks shows how NLP has moved towards and away from Wittgensteinian ideas since the 1960's. In these two papers, Wilks puts Wittgenstein in conversation with the history of NLP and contemporary Big Data methods. He explores Wittgenstein's influence on (the few) NLP researchers who studied him and demonstrates how contemporary projects in NLP and AI embody his ideas. In many ways, Wilks is the best case of a bad scenario. He does a fantastic job in nuancing Wittgenstein's views and I return to his work periodically throughout this section. However, as I argue at the end, his reading of Wittgenstein still falls short.

Wilks begins by describing how the emergence of Big Data has challenged the dominance of rule-based models in NLP, which were largely influenced by Chomsky's theories of language. On Wilks's view, Chomsky's attempt to derive a set of rules from language has proven rather impractical for NLP[57]. Wilks mentions one recent experiment which tried to find how many of Chomsky's "phase-structure rules" could be found in the "PennTree Bank," a ten-million word corpus used to develop many NLP technologies. The model produced *18,000* rules before the researchers decided to stop, as the number of rules kept increasing exponentially with the size of the corpus[58]. For Wilks, this experiment suggests that Chomsky was wrong to think language could be represented by a finite set of rules. While computers can create highly complex rule-based models, their complexity and size may be too cumbersome and complex to manage.

[55] Wilks, Yorick. "A Wittgensteinian computational linguistics?" 2008, p. 2.
[56] Wilks, Yorick. "Wittgenstein and Computational Linguistics." 2008, pg. 5.
[57] Wilks, Yorick. "Wittgenstein and Computational Linguistics," pg. 12.
[58] Wilks, Yorick. "Wittgenstein and Computational Linguistics," pg. 12.

Another limitation of "rule-based" models involves the pre-classification of words into databases. Take the "WordNet" database, for example[59]. WordNet has been used to train countless NLP technologies. It was created by dozens of Amazon Mechanical Turk employees who defined over 14 million words using hierarchical categories. However, WordNet has several problems. For one, these definitions often reflect racist and sexist stereotypes. For example, WordNet oddly and offensively defines the word "Hermaphrodite" as Person > Sensualist > Bisexual. Fixing this "bias," however, requires massive amounts of time and money. And even if these categories were to be "unbiased," the meaning of a word cannot always be captured by its classification in some hierarchy. These labels leave out important information about a word's context, use, and tone. Additionally, merely assigning words to categories and properties doesn't tell the computer how to navigate these classifications. As a result, rule-based models are ridden with "human bias," require massive amounts of time and money to create, and often don't accurately represent the meaning of a word.

Modern Big Data methods claim to solve some of the problems of traditional classification. In NLP, the widespread collection of "Big Data" has enabled bottom-up, self-adjusting models to predict a word's use or place in language, supposedly without much top-down guidance, human intervention, or theoretical assumptions. According to computer scientist pioneer Yoshua Bengio, Big Data methods like deep learning have "entirely contradicted Chomsky's rule-based theory language."[60] Or as Chris Anderson, the Editor of Chief of Wired

[59] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019. Accessed from https://www.excavating.ai/

[60] Goldhill, Olivia. "Google Translate is a manifestation of Wittgenstein's theory of language." *QZ.* February 13, 2019. Accessed from https://qz.com/1549212/google-translate-is-a-manifestation-of-wittgensteins-theory-of-language/

Magazine, describes Big Data: companies like Google and Amazon "don't have to settle for wrong models," in fact, "they don't have to settle for models at all."[61] As Anderson writes:

> The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all… Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity [...] the numbers speak for themselves.[62]

Anderson argues that applying complex mathematics to massive amounts of data surpasses top-town, rule-based algorithms in accuracy and functionality. According to Anderson, Big Data "forces us to view data mathematically first and establish a context for it later." [63] In other words, Big Data is the "end of theory" and the start of ground-up "empiricism."[64]

Of course, Wittgenstein could have predicted the limits of rule-based and Chomskyian linguistics long before the advent of deep learning. It's not hard to find evidence for Wittgenstein's aversion to rule-based theories, top-down models, and logic-based explanations for language in the *Philosophical Investigations.* In §81, Wittgenstein rejects the idea that there can be a finite set of rules for language[65]. In §23, he argues that the multiplicity and complexity of language cannot be captured in one theory or model[66]. In §133, he writes that the aim of his philosophy is not to "refine or complete the system of rules for the use of our words in unheard-of ways."[67] In §109, Wittgenstein says that, when doing philosophy,

[61] Anderson, Chris. "End of Theory: The Data Deluge Makes the Scientific Method Obsolete" *Wired Magazine.* 2008. Accessed from https://www.wired.com/2008/06/pb-theory
[62] Anderson, Chris. "End of Theory: The Data Deluge Makes the Scientific Method Obsolete." 2008.
[63] Anderson, Chris. "End of Theory: The Data Deluge Makes the Scientific Method Obsolete." 2008.
[64] Anderson, Chris. "End of Theory: The Data Deluge Makes the Scientific Method Obsolete." 2008.
[65] Wittgenstein, Ludwig. *Philosophical Investigations.* §81.
[66] Wittgenstein, Ludwig. *Philosophical Investigations.* §23.
[67] Wittgenstein, Ludwig. *Philosophical Investigations.* §133.

> We may not advance any kind of theory. There must not be anything hypothetical in our considerations. We must do away with all explanation, and description alone must take its place[68].

Rather than trying to come up with theoretical rules which govern language, Wittgenstein seeks to describe how language is used in specific, real life examples.

We can see why Wittgenstein's call to "Don't think, but look!"[69] is especially appealing to those who, like Anderson, call Big Data the "end of theory." However, it's important to understand that Wittgenstein's aversion to rule-based models isn't an outright rejection of their usefulness. Wittgenstein emphasizes that rule-based models can serve a clarifying purpose so long as they serve as an "object of comparison" (§131) rather than something which aims to represent the entire world[70]. Neither does Wittgenstein reject the idea that words can be assigned definitions and labels, like we might see in WordNet; he acknowledges that, sometimes, "naming something is rather like attaching a name tag to a thing" (§15)[71]. However, Wittgenstein points out that "defining words" is only one of many ways our words have meaning.

The ways in which Wittgenstein characterizes the relationship between rule-based models and his later ideas of language can help us understand the relationship between old rule-based models and new data-driven methods. One way to understand the transition of NLP into Big Data is by looking at Wittgenstein's early work in comparison to his later work. In *Tractatus Logico-Philosophicus,* Wittgenstein sought to ground language and thought in logic and representation. His "picture" theory of language held that words represented objects in the world

[68] Wittgenstein, Ludwig. *Philosophical Investigations.* §109.
[69] Wittgenstein, Ludwig. *Philosophical Investigations.* §66.
[70] Wittgenstein, Ludwig. *Philosophical Investigations.* §131.
[71] Wittgenstein, Ludwig. *Philosophical Investigations.* §15.

and that language is "a logical picture of facts."[72] As we saw, the later Wittgenstein doesn't so much reject this picture as re-contextualize his picture theory of language (while at the same time attempting to break out of its dogmatic tendencies.) As philosopher Anat Matar describes the transition from *Tractatus* to the *Philosophical Investigations*, Wittgenstein makes "the move from the realm of logic to that of ordinary language as the center of the philosopher's attention."[73] Interestingly, NLP practitioners seem to have a strong grasp of this relationship between the *Tractatus a*nd the *Investigations*[74]. Rather than see the *Investigations* as an argument against the *Tractatus*, NLP practitioners use the "picture theory" as a starting point for further development.

For example, the NLP firm Alegion has stated that Wittgenstein's early work in the *Tractatus Logico-Philosophicus* can help explain how preprocessing and pre-classifying data connects to machine learning methods. In "Understanding The Challenges Of NLP Through Wittgenstein," [75] (a blog post apparently written by an unnamed employee of Alegion), the unnamed author argues that the *Tractatus*'s description of language as "a logical picture of facts" is exactly what one does "when preprocessing data for NLP models."[76] As the Alegion author writes:

> For example, when I say "A willow tree sways in the wind," my words should invoke a concrete picture of the scene. If we were to take out all of the stopwords, or common words, we would be left with "willow", "tree", "sways", "wind" and still be able to get the gist of the meaning. This is exactly what we do when preprocessing data for NLP models.[77]

[72] Matar, Anat. "Ludwig Wittgenstein." *The Stanford Encyclopedia of Philosophy*. 2018. Accessed from https://plato.stanford.edu/archives/sum2018/entries/wittgenstein
[73] Matar, Anat. "Ludwig Wittgenstein." *The Stanford Encyclopedia of Philosophy*. 2018.
[74] I can't help but wonder in what other ways Wittgenstein's ideas are experienced and embodied by the act of building practical improvements on rule-based algorithms.
[75] "Understanding The Challenges Of NLP Through Wittgenstein." *Alegion.* Accessed from https://content.alegion.com/blog/nlp-part-ii-understanding-the-challenges-of-nlp-through-wittgenstein
[76] "Understanding The Challenges Of NLP Through Wittgenstein." *Alegion.*
[77] "Understanding The Challenges Of NLP Through Wittgenstein." *Alegion.*

The blogger argues that pre-processing steps such as removing stop words and normalizing the text are attempts to reduce language to a "picture of facts." However, the Alegion blogger claims that Wittgenstein's work can help NLP practitioners see that language is more than stating facts about the world. As Wittgenstein later describes in the *Philosophical Investigations*, words do much more than represent facts about the world. In §23, Wittgenstein writes:

> But how many kinds of sentences are there? Say assertion, question, and command?—There are countless kinds: countless different kinds of use of what we call "symbols", "words", "sentences". And this multiplicity is not something fixed, given once for all; but new types of language, new language-games, as we may say, come into existence, and others become obsolete and get forgotten. (We can get a rough picture of this from the changes in mathematics.)[78]

According to the Alegion blogger, Wittgenstein's later work "evolves to better match the complexity of communication," particularly when Wittgenstein theorizes language as "a tool with which we play different games or patterns of intention."[79] According to the blogger, the *Investigations* mirrors how contemporary NLP has evolved from solely relying on pre-classified definitions to looking for patterns in language use.

Let's now consider a particularly influential example in which Wittgenstein's ideas have been used to propel practical improvements in NLP. In 2002, Google scientists credited massive improvements to their search engine to Wittgenstein's "theories about how words are defined by context."[80] Before 2002, Google's search engine was based on a synonym system. At that time, their algorithm worked like this: if people searched for "pictures of dogs" and "pictures of puppies" in the same session, the algorithm would learn that "dogs" and "puppies" were

[78] Wittgenstein, Ludwig. *Philosophical Investigations.* §23.
[79] "Understanding The Challenges Of NLP Through Wittgenstein." *Alegion.*
[80] Levy, Steven. "How Google's Algorithm Rules The World." *Wired Magazine.* February 2, 2010. Accessed from https://www.wired.com/2010/02/ff_google_algorithm/

synonyms. Or if people searched for “hot water” and “boiling water” interchangeably, the program would conclude that “boiling” and “hot” were similar words. In 2001, however, engineers at Google realized that their search engine was also concluding that “hot dog” meant the same thing as “boiling puppy.” They realized that their algorithm was defining words by their use in one situation - search engine activity - rather than a variety of other situations those words were embedded in[81].

In 2002, Google Fellow Amit Singhal fixed this problem with a supposedly Wittgensteinian-inspired solution[82]. In §340, Wittgenstein writes: “One cannot guess how a word functions. One has to look at its use and learn from that.”[83] Likewise, Singhal’s team built an algorithm to “crawl and archive” billions of billions of documents and Web pages to see how “hot dog” was used[84]. Eventually, Google engineers found that “hot dog” was used in situations which also contained words like “bread,” “mustard,” and “baseball games.” By looking at how words were used not only in the search engine, but in other “language-games,” Google’s algorithm learned to understand the meaning of the word beyond a simple association method. Today, if you Google “Gandhi bio,” the algorithm knows that “bio” means “biography”, while if you type “bio warfare,” the algorithm knows it means “biological.” In a sense, the algorithm has learned the meaning of the word by its use.

In the era of Big Data and machine learning, it should not be surprising that Wittgenstein’s notion of “meaning-as-use” is perhaps the most common idea taken up by practitioners of NLP. In §43 Wittgenstein writes: “For a large class of cases—though not for

81 Levy, Steven. “How Google’s Algorithm Rules The World.” *Wired Magazine.* February 2, 2010. Accessed from https://www.wired.com/2010/02/ff_google_algorithm/
82 Levy, Steven. “How Google’s Algorithm Rules The World.” February 2, 2010.
83 Wittgenstein, Ludwig. *Philosophical Investigations.* §340.
84 Levy, Steven. “How Google’s Algorithm Rules The World.” February 2, 2010.

all—in which we employ the word "meaning" it can be defined thus: the meaning of word is its use in the language."[85] An interpretation of this claim can arguably be seen in data-driven methods of NLP. For instance, in "NLP word representations and the Wittgensteinian philosophy of language," machine learning scientist Christian Perone argues that methods like word embeddings and Word2Vec are comparable to Wittgenstein's idea of meaning as use[86]. Perone compares word embeddings to §11, in which Wittgenstein describes words as tools in a tool-box. In §11, Wittgenstein writes:

> Think of the tools in a tool-box: there is a hammer, pliers, a saw, a screw-driver, a rule, a glue-pot, glue, nails and screws.—The functions of words are as diverse as the functions of these objects. (And in both cases there are similarities.)[87]

According to Perone, word-embeddings resemble this idea because they treat language like a "tool," in the sense that each number in a vector can represent a different "function" of that word. For example, an LSA vector can represent how one document uses words in comparison to other documents, while a Word2Vec vector can represent lots of information about a word use. Perone argues that Word2Vec is like Wittgenstein's "meaning as use" because it calculates the "sum of all its uses."[88]

Perone also argues that Distributional Hypothesis, the assumption underlying word-embeddings, is based on Wittgenstein's idea of use. Keep in mind that Perone interprets Wittgenstein's notion of a word's "use" as its "co-occurrence" with other words. Perone credits

[85] Wittgenstein, Ludwig. *Philosophical Investigations.* §43
[86] Perone, Christian. "NLP word representations and the Wittgenstein philosophy of language." May 2018. Accessed from http://blog.christianperone.com/2018/05/nlp-word-representations-and-the-wittgenstein-philosophy-of-language/
[87] Wittgenstein, Ludwig. *Philosophical Investigations.* §11
[88] Admittedly, it's hard to imagine Wittgenstein saying something like "the meaning of a word is the sum of all its uses." And to be fair, Perone also sees this. At the end of his paper, he admits that Word2Vec "seems to be converging to an approximation of the average meaning of a word instead of capturing the polysemy inherent in language," in contrast with how Wittgenstein emphasized the multiplicity of language. However, Perone thinks this convergence may be happening because each word is represented by one vector. On the other hand, I suspect it's because Perone's statistical interpretation of "use" is misleading. I will return to this later.

this interpretation to John Firth, who says that Wittgenstein's "meaning as use" means something like "you shall know a word by the company it keeps!" Similarly, Perone understands a word's "use" to mean "nearby words." This is how Perone argues that Wittgenstein's ideas align with the "distributional hypothesis," or the assumption that word vectors which occur near each other convey similar meaning.

A summary of the comparisons between Wittgenstein, Word2Vec, and the distributional hypothesis can be seen in "Neural Networks and Philosophy of Language: Why Wittgenstein's theories are the basis of all modern NLP," in which software engineer Massimo Belloni writes:

> And it's now quite clear where the Wittgenstein's theories jump in: context is crucial to learn the embeddings as it's crucial in his theories to attach meaning. In the same way as two words have similar meanings they will have similar representations (small distance in the N-dimensional space) just because they often appear in similar contexts. So "cat" and "dog" will end up having close vectors because they often appear in the same contexts: it's useful for the model to use for them similar embeddings because it's the most convenient thing it can do to have better performances in predicting the two words given their contexts.[89]

The final comparison between Wittgenstein and NLP I want to draw attention to is between his idea of "family resemblances" and unsupervised "clustering" methods. Let's start with "family resemblances." In the *Philosophical Investigations,* Wittgenstein challenges the notion that we can know the meaning of every word by giving its definition or appealing to its "essential" properties. When it came to the definition of "game," for example, Wittgenstein showed that there is no one definition or essential property shared by board-games, ball-games, card-games, Olympic games, and mind-games. In §66, Wittgenstein writes that different uses of the word "game" are connected by "similarities, relationships, and a whole series of them at

[89] Belloni, Massimo. "Neural Networks and Philosophy of Language: Why Wittgenstein's theories are the basis of all modern NLP." *Towards Data Science.* Accessed from https://towardsdatascience.com/neural-networks-and-philosophy-of-language-31c34c0796da

that."[90] In §67, Wittgenstein decides to call the networks of small and large similarities which connect words "family resemblances":

> I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way.—And I shall say: 'games' form a family.[91]

As Wilks points out, one intuitive way to understand Wittgenstein's notion of family resemblances is the idea that "one could observe a family's pictures and spot a special kind of family face, but without implying there was any set of features shared by all those with the 'family face.'"[92]

Wilks argues that a technical interpretation of Wittgenstein's family resemblances influenced NLP as early as the 1960's, when British computer scientist Karen Spark Jones and her colleagues were working on classification problems in automated Information Retrieval.[93] Rather than coming up with the "defining properties" for every object in a database, Jones and her colleagues were trying to develop a mathematical procedure which could create "worthwhile" classifications for objects[94]. Computer scientist Robert Needham described Jones' problem as the following:

> Given a set of objects, and a list of properties of each, to find procedures for grouping the objects into subsets the members of which have in a defined sense a mutual `family resemblance'. We are thus concerned here with the stage before the usual classification procedures….This is a sorting problem and so presents no more than technical difficulties. The problem is that of discovering, given a collection of objects, what would be a worthwhile classification for them and for similar collections that is, the problem of defining classes, not of using them once defined[95].

[90] Wittgenstein, Ludwig. *Philosophical Investigations.* §67
[91] Wittgenstein, Ludwig. *Philosophical Investigations.* §67
[92] Wilks, Yorick. "Wittgenstein and Computational Linguistics." 2008, p. 18.
[93] Wilks, Yorick. "Wittgenstein and Computational Linguistics." 2008, p. 19.
[94] Wilks, Yorick. "Wittgenstein and Computational Linguistics." 2008, p. 19.
[95] Wilks, Yorick. "Wittgenstein and Computational Linguistics." 2008, p. 17-19.

In turn, Jones developed an "unsupervised statistical clumping" theory for organizing thesaurus entries. According to Wilks, Jones' method relates to family resemblances in that it produces "classifications/clumps in the data whose members did not all share any single feature used in the classification."[96] Thus, Wittgenstein's family resemblances presents an alternative way of defining and classifying not only words in relation to one another, but also data.

Fifty years later, I think family resemblances can be seen in unsupervised machine learning "clustering" techniques. Recall that, rather than define a word by its classification in some hierarchy of properties, unsupervised machine learning algorithms predict the meaning of a word by looking at its data structure. One popular application of unsupervised machine learning is clustering or cluster analysis, in which a model is trained to classify data into groups, also known as clusters, classes, or buckets. There are two kinds of clustering techniques. In traditional "K-means clustering," each item can only belong to one cluster. In "fuzzy clustering," one data point can belong to multiple clusters and its level of belonging or "membership" can vary by degrees[97]. Typically, the goal of clustering is for items in the same cluster to be as similar as possible while items in different clusters are as dissimilar as possible. In each cluster, the structure of the data should share certain kinds of "similarity measures." These could include the distance within and between clusters, how the data points are connected, and how intensely are linked.

Some have linked these similarity measures to Wittgenstein's idea of family resemblances. For example, in their "Data clustering based on family resemblance," computer scientists Yu Xiao and Jian Yu develop a k-means clustering technique they claim is inspired by

96 Wilks, Yorick. "Wittgenstein and Computational Linguistics." 2008, p. 19.
97 "Fuzzy Clustering." *Wikipedia.* Accessed from https://en.wikipedia.org/wiki/Fuzzy_clustering

Wittgenstein's family resemblances[98]. Xiao and Yu use k-means clustering to perform object recognition in photographs. The key thing to know about object recognition is that images are made of pixels, and the pixels which represent a chair, human body, or a hand-written "z" look different than the paper around the "z" or the sky in the background. The challenge then is to build an algorithm which can differentiate between the pixels that belong to objects versus the ones that don't. Using k-means clustering, Xiao and Yu created an unsupervised algorithm which makes clusters corresponding to the shape of objects in photos. In their case, each pixel either belongs to a cluster, or it doesn't.

On Xiao and Yu's view, family resemblances "emphasizes that things in the same concept are connected by overlapping sets of features, but not necessarily a common set of features."[99] They describe their "Wittgensteinian" idea of clusters in the following way:

> ...any two things in the same concept may not be similar to each other but must have a similarity connected path between them, in which any two directly linked things are highly similar to each other. Therefore, if a cluster is meaningful in perception, a cluster should represent a concept.[100]

Thus, rather than make clusters based on the similarities of the pixels themselves - for example, saying all dark pixels should be clustered with other dark pixels - Xiao and Yu decide to build clusters according to their "similarity connected path." If I follow Xiao and Yu correctly, a group of pixels becomes a meaningful "concept" or a recognizable "object" when it has a strong connection between its data points. Two data points in the same cluster need to be connected by

[98] Xiao, Yu and Yu, Jian. "Data clustering based on family resemblance." *2010 3rd International Congress on Image and Signal Processing.* Yantai, 2010, pp. 1373-1377. Accessed from https://ieeexplore.ieee.org/document/5648220.
[99] Xiao, Yu and Yu, Jian. "Data clustering based on family resemblance." *2010 3rd International Congress on Image and Signal Processing.* Yantai, 2010, pp. 1373-1377. Accessed from https://ieeexplore.ieee.org/document/5648220.
[100] Xiao, Yu and Yu, Jian. "Data clustering based on family resemblance." pp. 1373-1377.

a strong path *and* the data located alongside that path also needs to be connected to each other. While they don't explicitly draw this connection, Xiao and Yu's definition of a meaningful "concept" reminds me of Wittgenstein when he says that creating family resemblances is like spinning a thread by twisting fibres together, and that "the strength of the thread does not reside in the fact that some one fibre runs through its whole length, but in the overlapping of many fibres." (§66).[101]

Admittedly, I'm not entirely clear on how the mathematics and computer science behind Xiao and Yu's clustering technique relate to family resemblances. However, their paper inadvertently demonstrates another utterly Wittgensteinian point: in deciding what should count as a cluster and how to draw boundaries between clusters, computer scientists exercise their judgment at every step. In k-means clustering, for instance, since each data point can only belong to one cluster, humans have to use their judgment to decide what to do with data points caught in the middle. For example, Xiao and Yu deal with pixels caught between objects by setting a minimum criteria for the strength of their "similarity-path" connections[102]. Thus, how to differentiate between clusters, and thus concepts they represent, requires careful human judgement.

In computer science, however, pointing out the role of human judgment can be akin to calling something "arbitrary," "subjective," or "biased." Couldn't one say that Xiao and Yu's algorithm is just as arbitrary as the old fashioned way of classifying objects by hand? Well, yes - according to the authors of *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*[103]. When describing how k-means clustering classifies the overlapping pixels caught

[101] Wittgenstein, Ludwig. *Philosophical Investigations.* §66.
[102] Xiao, Yu and Yu, Jian. "Data clustering based on family resemblance." pp. 1373-1377.
[103] Bezdek, J.C., Keller, J., Krisnapuram, R., Pal, N. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Springer. 1999.*

between the “edges” of objects in photos, Bezdeck et. al write that “the definition of what constitutes an edge is rather vague, heuristic, and even subjective."[104] Thus, to “remove” human subjectivity and to lower the impact of wrong “subjective” decisions, the authors suggest a more “automated” and flexible method of clustering objects: fuzzy clustering. As I will show, fuzzy clustering can also be understood using Wittgensteinian ideas regarding family resemblances and “blurry concepts.”

First, some more background: Fuzzy clustering is an unsupervised learning method in which one data point can belong to multiple clusters in varying degrees[105]. With fuzzy clustering, each cluster is given a “centroid” or representative example[106], and the farther away some data point is from that centroid, the less likely it is part of that cluster, and the more likely it is in another cluster. These low-probability outliers are called “fuzzy edges,” and they help account for the fuzziness of real life, such as shadows or blurs in pictures. In turn, this method lessens the effect of having the “wrong” definition for “edge.” That is, instead of a 100% of the data point belonging to the wrong cluster, there is only a 50% or 20% chance of it being assigned to the wrong cluster.

However, overlapping data points seem to pose a problem for edge detection: if a data point can belong to more than one cluster, isn’t that cluster less precise? In other words, how can we use fuzzy clusters to draw sharp boundaries around objects? These questions were raised in a fascinating StackExchange forum titled “On Wittgenstein's family resemblance and machine learning,” an online discussion between (apparently) a mix of computer scientists and people

[104] Bezdek, J.C., Keller, J., Krisnapuram, R., Pal, N. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Springer. 1999.*
[105] Fuzzy Clustering” *Wikipedia.* Accessed from https://en.wikipedia.org/wiki/Fuzzy_clustering
[106] In this sense, we can link these representative “centroid” examples to Wilks’s idea of the representative face of a family.

interested in philosophy[107]. In this forum, one commenter expressed skepticism that fuzzy clusters could be used to draw sharp edges. Just like the computer scientists who felt that the definition of an edge in k-means clustering was "rather vague, heuristic, and even subjective," this commenter wasn't sure that having "blurrier concepts" would be a solution. But as another commenter pointed out[108], Wittgenstein's defense of family resemblances more or less resembles the reasons why fuzzy clustering is used and why it is particularly successful in edge detection.

In §71, Wittgenstein writes that we might be tempted to think the idea of family resemblances has rendered "game" into an unclear concept with "blurred edges."[109] But then Wittgenstein asks, "Is it even always an advantage to replace an indistinct picture by a sharp one?"[110] This isn't a normative question, but a descriptive one: in what cases do blurry concepts serve a useful purpose? In §76-78, Wittgenstein gives an answer to this descriptive question. In these passages, he describes what it means to draw sharply defined boundaries over an indefinite region. Wittgenstein points out that there are many ways to draw boundaries which "correspond" to some indefinite space, but the one we choose to draw depends on our purpose. In §77, he writes:

> For imagine having to sketch a sharply defined picture 'corresponding' to a blurred one. In the latter there is a blurred red rectangle: for it you put down a sharply defined one. Of course-several such sharply defined rectangles can be drawn to correspond to the indefinite one. -But if the colours in the original merge without a hint of any outline won't it become a hopeless task to draw a sharp picture corresponding to the blurred one? Won't you then have to say: "Here I might just as well draw a circle or heart as a rectangle, for all the colours merge.

[107] "On Wittgenstein's family resemblance and machine learning." *Stack Exchange Forum: Philosophy.* January 26, 2017. Accessed from https://philosophy.stackexchange.com/questions/40581/on-wittgensteins-family-resemblance-and-machine-learning

[108] "On Wittgenstein's family resemblance and machine learning." *Stack Exchange Forum: Philosophy.* January 26, 2017.

[109] Wittgenstein, Ludwig. *Philosophical Investigations.* §71

[110] Wittgenstein, Ludwig. *Philosophical Investigations.* §71

> Anything-and nothing-is right." And this is the position you are in if you look for definitions corresponding to our concepts in aesthetics or ethics.
>
> In such a difficulty always ask yourself: How did we learn the meaning of this word ("good" for instance)? From what sort of examples? in what language-games?[111]

In the first line, Wittgenstein asks us to imagine that we have to sketch a sharp picture which "corresponds" to our blurred one -- imagine this is the situation we are in when doing edge detection. The scare quotes suggest that it is not Wittgenstein who wants to draw sharp boundaries for the sake of doing so. In §76, Wittgenstein writes: "If someone were to draw a sharp boundary I could not acknowledge it as the one that I too always wanted to draw, or had drawn in my mind. For I did not want to draw one at all. His concept can then be said to be not the same as mine, but akin to it."[112] At the same time, it isn't necessarily a *problem* if someone else (for example, an algorithm) draws a sharp picture -- Wittgenstein can appreciate the similarities between a sharpened picture and blurred one just as he can recognize the differences. We must be careful not to misinterpret Wittgenstein's point here. We shouldn't conclude that all sharp boundaries are *arbitrary* just because many sharp boundaries can be "akin" to the blurry concept. Wittgenstein raises this concern through the voice of a skeptical interlocutor: "'Here I might just as well draw a circle or heart as a rectangle, for all the colours merge. Anything-and nothing-is right.'"[113] Wittgenstein notes that this sense of relativity is also present when we "look for definitions corresponding to our concepts in aesthetics or ethics."[114] That is, it is easy to think our aesthetic and moral definitions are completely relative just because there are many different uses of the words "goodness" or "beauty." In response to this skepticism, Wittgenstein suggests

111 Wittgenstein, Ludwig. *Philosophical Investigations.* §77.
112 Wittgenstein, Ludwig. *Philosophical Investigations.* §76.
113 Wittgenstein, Ludwig. *Philosophical Investigations.* §77.
114 Wittgenstein, Ludwig. *Philosophical Investigations.* §77.

that we look closely at how our ethical and aesthetic concepts are actually used in our world: “In such a difficulty always ask yourself: How did we learn the meaning of this word ("good" for instance)? From what sort of examples? in what language-games?”[115] The cure to relativism is not to theorize even more abstractly, but to look at how concepts are actually used and differentiated from each other.

Let’s now put Wittgenstein in conversation with fuzzy clustering. Say we have been asked to develop a definition for an edge for object recognition and we can’t help but think about the many different ways a computer scientist could define an edge. When faced with different “subjective” definitions for an edge, it is tempting to buckle down and look even harder for an ultimate and essential definition of an “edge.” However, Wittgenstein tells us that when we are feeling this sense of relativity, we need to ask ourselves how we learned about a concept in the first place and to look at real life examples. In a sense, fuzzy clustering methods do just that. We use unsupervised training so the algorithm can “learn” the meaning of an edge by looking at real life examples of data. Furthermore, by allowing data points to have partial membership in different clusters, fuzzy clustering accounts for the fact that objects in real life have edges that blur into shadows, other objects, and “noise” in the data – or Wittgenstein describes, fuzzy clustering shows how one data point can have “a family of meanings.” Just like family resemblances, fuzzy clustering treats the “overlapping” similarities between objects as a strength, not a weakness. In fact, as fuzzy clusters are proving to be better at edge detection than k-means methods, Wittgenstein’s contention that it is not “always an advantage to replace an indistinct picture by a sharp one” also rings true[116]. As one commenter in the fuzzy clustering and Wittgenstein StackExchange forum put it, even the sharpest lines “rely on multiple resemblances

[115] Wittgenstein, Ludwig. *Philosophical Investigations.* §77.
[116] Wittgenstein, Ludwig. *Philosophical Investigations.* §77.

and similarity measures, involve arbitrary choices and produce blurred results."[117] Fuzzy clusters, as an improvement over k-means clusters, more or less mirrors Wittgenstein's defense of blurry concepts over unnecessarily precise ones.

In summary, what does Wittgenstein share in common with NLP methods in the age of Big Data and machine learning? First, they prioritize real-life examples over theoretical or abstract theories of language. Second, they both show how rule-based models based on abstract theories of language can be reductive, overgeneralizing, and inaccurate. Third, they both claim to define the meaning of a word by its use. Fourth, they both try to incorporate the natural fuzziness of real day life into their concepts. And finally, the ongoing transition from relying on rule-based models to data-driven algorithms supposedly parallels Wittgenstein's movement from the *Tractatus* to the *Philosophical Investigations.* As I have come to discover, however, these comparisons have serious shortcomings.

NLP and Wittgenstein: Differences

While Big Data and Wittgenstein may share an aversion to theory and rule-based methods, this doesn't mean Wittgenstein automatically embraces the empiricism of Big Data. As Wilks points out, the call to "look at language data" is an "authentic Wittgensteinian demand," but Wittgenstein and NLP look at language data in very different ways[118]. Wittgenstein uses specific examples to examine the details of a particular situation. In contrast, NLP uses language data to generalize about language. Thus, Wilks doesn't believe that Wittgenstein would identify with the

[117] "On Wittgenstein's family resemblance and machine learning." *Stack Exchange Forum: Philosophy.* January 26, 2017.
[118] Wilks, Yorick. "A Wittgensteinian computational linguistics?" 2008, p. 2.

empiricism of Big Data. As a result, Wilks is wary of placing Wittgenstein "somehow closer to the anthropological-empirical tradition than to Chomsky."[119]

If Wittgenstein is not an empiricist, then what do we make of the comparisons between Wittgenstein and empiricist NLP methods? Take, for instance, Perone's argument that Wittgenstein's "meaning-as-use" is like "count-based methods based on co-occurrence" because they learn the "meaning of a word by the company it keeps," as John Firth puts it[120]. Or software engineer Massimo Belloni's claim that we only need to tweak "Wittgenstein's theories a bit" to say that "dogs are similar to cats because they often appear in the same contexts."[121] To evaluate these comparisons, we need to inspect the assumptions being made. First, these comparisons equate Wittgenstein's notion of a word's "use" with a word's "context." Secondly, they assume that a word's "context" can be represented by a data set. Third, they assume that a word's "use" can be represented by how data points interact with each other in some data set. This is how a word's "use" can be quantified as some mathematical property, such as a co-occurrence with other words, or a Term-Document matrix, or a machine learning generated prediction.

Let's focus on the second assumption first, that a word's "context" can be represented by a dataset. This assumption poses a major technical problem for Wilks. Wilks argues that most datasets in NLP (like WordNet, for example) are too preprocessed and too small to accurately represent the real context of a word. For example, the pre-processing step of removing stop words like "the." "and," "or" can change the meaning of a phrase. (For example, the meaning of

[119] Wilks, Yorick. "A Wittgensteinian computational linguistics?" 2008, p. 2.
[120] Perone, Christian. "NLP word representations and the Wittgenstein philosophy of language." May 2018. Accessed from http://blog.christianperone.com/2018/05/nlp-word-representations-and-the-wittgenstein-philosophy-of-language/
[121] Belloni, Massimo. "Neural Networks and Philosophy of Language: Why Wittgenstein's theories are the basis of all modern NLP." *Towards Data Science*. January 7, 2019. Accessed from https://towardsdatascience.com/neural-networks-and-philosophy-of-language-31c34c0796da

the sentence "I didn't say she was *the* leader, but *a* leader," is not captured by "didn't say she was leader, but leader.") Likewise, small datasets cannot grasp the meaning of a word because they exclude possible contexts for that word. For example, a machine learning model trained on data from the healthcare industry may not understand how words are used in Amazon customer reviews. Wilks summarizes these problems in this way: "If meaning comes from… the company words keep, (as in Firth's much quoted phrase) then it changes inevitably as the *surrounding vocabulary* changes. Or as we might now put it in NLP, toy language experiments[122] do not scale up!"[123] For these reasons, Wilks thinks traditional datasets are not Wittgensteinian because they change the "surrounding vocabulary" which gives a word meaning[124]. Thus, a word's real "context" cannot be represented by datasets that do not capture natural, unstructured language.

On this view, however, if a dataset *were* to be unstructured and unprocessed, then it would capture a word's context and thus its meaning. Wilks thinks that naturally occurring conversations drawn from the World Wide Web - which are for all practical purposes *not* finite or pre-processed - can "appeal to use... in a very satisfying way."[125] He argues that this unstructured data is able to ground the meaning of words because they capture their "real usage."[126] In other words, because the internet hosts "real life" conversations - emails, Twitter threads, YouTube comments - it can get at a word's "real" use in a way that ordered ontologies (such as WordNet) cannot. Thus, for Wilks, a word's "use" *can* be captured by a dataset if the dataset includes the full "context" of that word's use. For these reasons, Wilks thinks that datasets which use unstructured data and methods which maintain the contextual relationship

[122] For reference: a "toy language" is a miniature computer processing language without the full capabilities of higher end processing language, often used as a proof of concept or prototype. We could think of Wittgenstein's 'block and slab' examples as a toy language in §2.
[123] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 20.
[124] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 20.
[125] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 20.
[126] Wilks, Yorick. "A Wittgensteinian computational linguistics?" 2008, p. 5.

between words – for example, word embeddings and Word2Vec – could be considered Wittgensteinian.

However, I have several problems with Wilks's conclusion. First of all, I don't think that Wittgenstein's "meaning as use" is captured by a word's immediately surrounding vocabulary. As philosopher Rob Chodat argues[127], Wittgenstein consistently emphasized that a word's context extended beyond its place in a sentence or "proposition." A proposition is also part of a "larger linguistic act," a "particular practice," a "socio-historical moment," and a "Lebensform," or form of life[128]. Chodat describes Wittgenstein's views on words and context in the following way:

> Not only do specific words mean in the context of a proposition, but a proposition means only as part of a larger linguistic act, a linguistic act means only in the context of a particular practice, a practice only in the context of a socio-historical moment, a socio-historical moment only in the context of a wider *Lebensform* [form of life.][129]

For Wittgenstein, understanding a word's context as its place in a sentence ignores all the other factors which give language meaning, such as tone, body language, and cultural norms. For Wilks, however, defining a word's context as its "immediately surrounding vocabulary" is an acceptable interpretation of Wittgenstein so long as these surrounding words are unprocessed, natural, and informative.

I also don't think that "context" captures Wittgenstein's "meaning as use." When we use language, our words don't get meaning because we are throwing them out into some context (as Word2Vec assumes.) Instead, they get meaning because we *do* something with them. That is,

[127] Chodat, Rob. "Appreciating Material: Criticism, Science, and the Very Idea of Method." *Ludwig Wittgenstein and Literature Colloquium,* 2019, p. 19.
[128] Chodat, Rob. "Appreciating Material: Criticism, Science, and the Very Idea of Method." *Ludwig Wittgenstein and Literature Colloquium,* 2019, p. 19.

when we make promises or argue or joke around, we are actively participating in shared activities with other speakers and listeners. Wittgenstein's concepts of "language-games" and "forms of life" are meant to differentiate from the notion of context. When Wittgenstein says the meaning of *most* words are in their use, he uses the word *language-game* to "bring into prominence the fact that the speaking of language is part of an activity, or of a form of life (§23)[130]. By form of life, Wittgenstein means something like "both our cultural practices *and* their connections to the natural conditions of our life," as philosopher Toril Moi puts it[131]. Learning a language, then, is "not to learn a set of names, but to be trained in -- to learn to recognize and participate in -- a vast number of human practices."[132] Wittgenstein offers some examples of language-games such as "Making up a story; and reading it," "Presenting the results of an experiment in tables and diagrams," "Play-acting," "Solving a problem in practical arithmetic," "Translating from one language into another," "Asking, thanking, cursing, greeting, praying."[133] To successfully play all of these language games, we have to understand how to participate in the shared social practices, or forms of life, in which they are embedded. These social practices are not only the routines or habits of a certain community, but also the background conditions which make these practices possible, such as the nature of our physical bodies, our histories, and even our institutions. All of this considered, I don't get "context" from forms of life and language games so much as *circumstance* or even *situation* – thus, I do not think Wittgenstein's "use" cannot be reduced to "surrounding vocabulary" or "context."

Still, this raises a question: can NLP models get at language games and forms of life? Is there a Big Data method which can include the sarcasm of an insult, the gender dynamics in a

[130] Wittgenstein, Ludwig. *Philosophical Investigations.* §23.
[131] Moi, Toril. *Revolution of the Ordinary.* University of Chicago Press, 2017, p. 55.
[132] Moi, Toril. *Revolution of the Ordinary.* University of Chicago Press, 2017, p. 55.
[133] Wittgenstein, Ludwig. *Philosophical Investigations.* §23.

flirty text, or the racialized history of a word? At first glance, it seems like no. Wittgenstein's emphasis on *participating* in language seems to suggest that the only way a machine could understand the meaning of, say, an invitation to a party, would be to actually attend parties. And while there is much more to be said about what it means for a machine to "understand" something, I tend to think that machines have to *participate* in language games and forms of life to grasp meaning. The *nature* of this participation, however, depends on how we characterize the relationship between NLP computational models and natural language, and thus between technologies and human life.

Wilks, for example, believes that NLP datasets are "formal objects, kept apart from language and its vagaries, and with only the meanings assigned to them by scientists."[134] Wilks describes both datasets and models as "metadescriptions" of our language[135]. By "metadescriptions," Wilks means "annotations" or "semantic features, types or markers, or logical descriptions" which are added to a text[136]. These models feature "language-like items," but they are distinct from natural language itself. As metadescriptions, they are "micro languages" which "do not have all the features of a full natural language" but can still be used to help clarify our language use[137]. Thus, NLP models cannot get at language games, but only represent them. At their best, these models are highly representative - like models built from the World Wide Web, which Wilks thinks can produce "justifiable meanings" because they are "linked directly to language corpora, chiefly by being built automatically from them."[138]

[134] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 31.
[135] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 9.
[136] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 9.
[137] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 9.
[138] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 31.

According to Wilks, NLP models are not part of natural language, nor are they translations of our language into mathematical terms. However, Wilks understands that Wittgenstein would have a problem with suggesting that metadescriptions lie *outside* of our natural language, in "some special space" of logic and semantics rather than "in the space of words."[139] At the same time, Wilks thinks it is difficult to make the case that "formal and semantic/linguistic languages are all miniature, functional, natural languages."[140] For Wilks, the idea that metadescriptions are part of our natural language amounts to the claim that "metadescriptions are no more than the translation of one language into another."[141] If this were true, then formalized computer language and mathematical formulas could all be replaced, one by one, by words in English. In fact, this was attempted by McDermott, who argued that formulas could be replaced by "computer-generated gensyms like G110004467."[142] However, Wilks calls this the "Gensym fallacy" and argues that "humans could *not* in fact manipulate such substituted forms unless they learned that language fully (so as to be "inside" it, as it were)."[143] Indeed, it is hard to imagine how we could "be inside" the language of WordNet databases, Term-Document matrixes, high-dimensional vectors, or learned linear transformations. Even Wittgenstein makes a similar claim in §513:

> Consider the following form of expression: 'The number of pages in my book is equal to a root of the equation $x^3 + zx — 3 = 0$.' Or: 'I have n friends and '$2n + 2 = 0$'. Does this sentence make sense? This cannot be seen immediately. This example shows how it is that something can look like a sentence which we understand, and yet yield no sense.[144]

[139] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 9.
[140] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 10.
[141] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 10.
[142] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 20.
[143] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 20.
[144] Wittgenstein, Ludwig. *Philosophical Investigations.* §513.

Furthermore, even if we could "get inside" and use mathematical language, Wilks thinks that "the risk of the symbols regaining ambiguity would return as in any language."[145] That is, if different communities start using mathematical notations in their language, Wilks argues that this would make it harder for scientists and mathematicians to "protect" the clarity and preciseness which is needed to accurately model language for NLP[146].

Returning to the question of whether NLP models can capture language games, then, is clearly a question about what we *do* with NLP models. As we saw, Wilks looks at models as representations of language. And while models are certainly used to "represent" reality, we must be careful to remember that representation" is a "crude" notion, as Moi calls it, especially "when it is used to cover all the things we do with words" and, as I argue, all the things we do with models[147]. While we can think of datasets and models as descriptions or representations, we must not think they are "idle" pictures on the wall (§291). If we think of a "machine drawing" as a "word-picture of the facts," then we might be misled into thinking of NLP models as passive descriptions outside of our language. As Wittgenstein writes:

> What we call "descriptions" are instruments for particular uses. Think of a machine-drawing, a cross-section, an elevation with measurements, which an engineer has before him. Thinking of a description as a word-picture of the facts has something misleading about it: one tends to think only of such pictures as hang on our walls: which seem simply to portray how a thing looks, what it is like. (These pictures are as it were idle)[148].

I think that Wilks falls into this trap, for example, when he characterizes the fact that mathematical terms change their meaning with human use as a problem of ambiguity -- ambiguity is only a problem if there is a pre-existing requirement for computational models to be

[145] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 10.
[146] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 24.
[147] Moi, Toril. *Revolution of the Ordinary.* University of Chicago Press, 2017, p. 13-14.
[148] Wittgenstein, Ludwig. *Philosophical Investigations.* §291.

a perfect representation of the real world - to idly hang on the wall like a mirror. Similarly, the fact that Wilks turns towards the internet for more "representative" and "real" language data reflects the expectation that "representation" is the only way for models to have "real" meaning. The problem with solely focusing on the representative aspect of models, however, is that it can make us lose sight of all the other things we *do* with these models and all the other ways they connect to the "real world.

While I agree with Wilks that datasets and models aren't in themselves natural language, I don't think they have to be "translated" into natural language to consider them part of our world of language. That is, we don't have to claim we speak *in* computer language to focus on what we *already* say *about* computers, in front of computers, and while using computers. NLP models are *already* part of our natural language because language is more than throwing words into the air -- it is an activity we participate in. The very act of creating, using, or applying an NLP model comes with its own language games and forms of life. For instance: "deciding how to pre-process a dataset," "presenting an NLP model," "describing an NLP method," "writing code with a team of computer scientists," "hiring an Amazon Turk employee to label words," or "downloading user activity." Just because we can't speak in computer language does not mean that computer models are isolated from our language games or otherwise "protected" from the "changes and vagueness of real words in use by human beings,"[149] as Wilks puts it. As any person who studies "bias" in computer systems will tell you, the human *always* enters the machine. Our forms of life are deeply embedded in NLP datasets.

Ultimately, Wilk's description of mathematical models and datasets as "finite" and "isolated" from our natural language sets him up for the "problem" of figuring out how to make

[149] Wilks, Yorick. "Wittgenstein and computational linguistics." 2008, p. 39.

NLP "represent" natural language. Instead of using Wittgenstein to investigate how NLP models are *already* connected to the world of language, Wilks uses Wittgenstein to bridge the "gap" between technology and the world. For these reasons, I don't think Wilks succeeds in arguing for a "Wittgensteinian computational linguistics." However, I do think this separation between NLP models and natural language is not a problem unique to Wilks. Instead, I think it is part of a larger phenomenon – a picture of technology in which technology is separate from the human world. In the next chapter, I explore this picture at length. I show why the notion that computer models are isolated from natural language is not just a technical issue, but also an ethical one.

Chapter 3

> It is not the business of philosophy to resolve a contradiction by means of a mathematical or logico-mathematical discovery, but to render surveyable the state of mathematics that troubles us[150].
>
> (Wittgenstein, Ludwig. *Philosophical Investigations.* §125.)

In this chapter, it is no longer my purpose to use Wittgenstein as a way to resolve technical problems in NLP or AI. Instead, I explore questions related to the ethics of technology and AI. It may seem strange to jump from a technical perspective on Wittgenstein to an ethical one, but the ethical and technical are intimately connected for Wittgenstein. Trained as an engineer and mathematician, Wittgenstein wrote the *Philosophical Investigations* just as Europe was being ravaged by all sorts of new technologies, from racial classification systems, to death camps, to atomic weapons. Wittgenstein's later philosophy is extremely relevant for understanding our own technological revolution. As Juliet Floyd stresses, Wittgenstein's work:

[150] Wittgenstein, Ludwig. *Philosophical Investigations.* §125.

> grapples with what it is to be human in the midst of a computational and technological revolution that is rapidly reshaping our manners of conducting and expressing social and personal relationships and ties as well as our relationship to the earth and our biologies.

I use Wittgenstein's methods to clarify my problems with how we speak and write about the "ethics of technology" and "Artificial Intelligence" or "AI." By "Wittgenstein's methods," I mean working with examples, engaging in a "grammatical investigation" of how our language is used, looking at what opposing views hold in common, and situating the ethics of technology and AI in everyday life. I use these methods to describe how we are "held captive" by a "picture" of technology, to use Wittgenstein's words (§115,) in which technology and the human/social/ethical belong to two completely separate worlds[151]. Using a series of examples, I show how this picture leads to confusions in the way we talk and write about the ethics of technology and Artificial Intelligence.

Let's take one confusion at a time. When I say I have a problem with the "ethics of technology," I am not only talking about the obviously wrong claim that technology is "objective" or "neutral." Although this claim posits technology as something *outside* of human subjectivity, judgement, and ethical claims, this argument has already been addressed by a wide range of scholarship. Instead, I am more concerned with how said scholarship argues against this claim. Take, for example, Kate Crawford and Trevor Paglen's critique of the database ImageNet as "subjective," "biased," and "socially constructed." By adhering to the objective/subjective and essentialist/socially constructed binaries, this type of criticism of technology fails to address the reasons *why* these binaries are harmful in the first place. For similar reasons, when some critics of technology spend their efforts trying to show how human "bias" enters the machine, they fail

151 Wittgenstein, Ludwig. *Philosophical Investigations.* §115.

to consider whether *bias* is really what is causing their problems. To bring this out, I use a Twitter conversation which debates whether or not algorithms can be "biased." As I show, both of these examples about the so-called "ethics of technology" is so obsessed with explaining "human-machine interactions" that they forgets about "human-human interactions in the presence of machines," to paraphrase Juliet Floyd[152]. By focusing on overcoming the theoretical separation between the world of technology and the world of humans, ethics, and society, they have already made the presumption that there is a *gap* between our technology and our human world which must be explained.

At the same time, I also have a problem with the claim there is *no* meaningful separation between technology and humans, an argument made by some AI enthusiasts, philosophers, and post humanist scholars alike. I find that the anthropomorphizing of technology leads to theoretical debates about "general artificial intelligence," "robot rights," and whether "machines can think," questions which are so abstracted and focused on a nonexistent technologies in the future that we once again forget about how *human* intelligence and *human* rights are affected by the AI systems which *already* exist. However, I am also intrigued by the entirely different argument made by some post-humanists who claim that the machine/ human distinction has been effectively compromised by the history of chattel slavery and the objectification of the Black subject. But while I think technological development *must* be situated in the acceleration of the economy by slavery and anti-Blackness, I am not sure what to make of the claim that Black humans *became* part of the concept of "machine" to show this. As I will show, these different conversations about the ethics of technology and AI hold more in common than we might think.

152

In all cases, it seems that the way we talk about technology seems to get in the way of having clear conversations about the *ethics* of technology.

Example 1: ImageNet Roulette

ImageNet is a project which aims to "map out the entire world of objects," according to co-creator Fei-Fei Li[153]. ImageNet includes pictures of everything from fruits, cars, cities, furniture to lawyers, doctors, assistant professors, and associate professors, so long as the image can be represented by a noun and classified into a hierarchy inherited from ImageNet's parent database, WordNet. As of 2018, over fourteen million images have been collected, classified, and organized into the database[154]. These labeled images, collected from online websites often without the original owners' knowledge or consent, have been used to train countless machine learning algorithms for image recognition tasks, making ImageNet the so-called "Rosetta Stone" of the image recognition boom[155].

Despite being one of the most widely used pre-labeled training sets, only recently has ImageNet been publicly scrutinized. In 2019, ImageNet was exposed for its classification of images related to the "Person" category, with critics calling these classifications "racist," "sexist," and "just plain bizarre."[156] For every image, a group of two or three Amazon Mechanical Turk workers select which concepts should represent it. Each concept in ImageNet is organized into classes, categories, and subclasses. For example, the class "Adult Body" is classified under Natural Object > Body, and only has two corresponding concepts: "adult female

153 Crawford, Kate and Paglen, Trevor. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019. Accessed from https://www.excavating.ai/
154 Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019.
155 Bernal, Natasha and Dodds, Laurence. "'Rosetta Stone' of AI slammed for 'racist' classification of people's faces." *Telegraph.* 2019. Accessed from https://www.telegraph.co.uk/technology/2019/09/17/rosetta-stone-ai-slammed-racist-classification-peoples-faces/
156 Bernal and Dodds, 2019.

body" and "adult male body." Meanwhile, the concept "Hermaphrodite" is organized as Person > Sensualist > Bisexual. Anti-Black racial slurs are also especially prevalent. Other offensive concepts include Bad Person, Call Girl, Drug Addict, Closet Queen, Convict, Crazy, Failure, Flop, Fucker, Hypocrite, Jezebel, Kleptomaniac, Loser, Melancholic, Nonperson, Pervert, Prima Donna, Schizophrenic, Second-Rater, Spinster, Streetwalker, Stud, Tosser, Unskilled Person, Wanton, Waverer, and Wimp.

The labels in the "Person" category are what motivated Kate Crawford and Trevor Paglen to create "ImageNet Roulette," a publicly available image recognition software trained on ImageNet's "Person" images and labels. ImageNet Roulette works like this: you upload a picture of yourself and ImageNet Roulette's algorithm assigns your face a label, anything from "Associate Professor" to "Man" to "Slattern." As an online art project, ImageNet Roulette instantly went viral, giving people a rare glimpse into the labels which are quietly used to train image recognition technologies, as well as an opportunity to see how their own face might be labeled. When I uploaded my picture to ImageNet Roulette, I was shocked to see my face framed by a thin green box and succinctly labeled "Zionist." Shocked because, that same day, I had just returned home from protesting against former Israeli Minister of Justice Tzipi Livni on account of her war crimes against Palestinian civilians. To put it plainly: I consider myself to be vocally and actively anti-Zionist. But there I was, looking at my "Zionist" label and being given no justification or explanation. I was overwhelmed with uneasiness as I realized just how quickly and mysteriously my humanity had been inaccurately and offensively objectified by ImageNet - which is exactly the feeling the creators of ImageNet Roulette intended to invoke. According to Paglen and Crawford, ImageNet Roulette is an "object lesson" in "what happens when people

are categorized like objects."[157] ImageNet Roulette is a relatively easy way to learn this lesson, but in most cases, this lesson is learned the hard way.

In an essay explaining their project, Crawford and Paglen explore the politics involved in labeling images. Drawing from theory in aesthetics and philosophy, they point to several major problems with ImageNet. They spend most of their essay challenging ImageNet's assumption that there are "uncomplicated, self-evident, and measurable ties between images, referents, and labels."[158] Here is how they describe their problems with ImageNet:

> First, the underlying theoretical paradigm of the training sets assumes that concepts—whether "corn", "gender," "emotions," or "losers"—exist in the first place, and that those concepts are fixed, universal, and have some sort of transcendental grounding and internal consistency. Second, it assumes a fixed and universal correspondences between images and concepts, appearances and essences. What's more, it assumes uncomplicated, self-evident, and measurable ties between images, referents, and labels. In other words, it assumes that different concepts—whether "corn" or "kleptomaniacs"—have some kind of essence that unites each instance of them, and that that underlying essence expresses itself visually[159].

Crawford and Paglen take issue with the fact that ImageNet uses subjective labels which are disguised as objective descriptions of the image. Not only do these labels inaccurately represent the fluid reality at hand, but they reify offensive and exclusionary concepts.

To oppose the essentialism of ImageNet's labels, Crawford and Paglen take the well-trodden path of arguing that the meanings of images are "socially constructed." In their essay, they write that the relationship between label, image, and object is flexible and unstable, and how

[157] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019.
[158] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019.
[159] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019.

one interprets an image is heavily dependent on one's personal views and "cultural context."[160] As Crawford and Paglen put it:

> The circuit between image, label, and referent is flexible and can be reconstructed in any number of ways to do different kinds of work. What's more, those circuits can change over time as the cultural context of an image shifts, and can mean different things depending on who looks, and where they are located. Images are open to interpretation and reinterpretation[161].

Crawford and Paglen argue that the "very unstable" [162] relationship between images and reality is further destabilized by technologies like ImageNet, who take these images even further outside of their original context[163]. According to them, the moment these images are removed from the real world is the moment "when things get strange."[164] The image is "strange" because the image, as a representation, doesn't correspond to reality. Likewise, the labels are "strange" because the label, as a representation, doesn't correctly correspond to the image. As a result, ImageNet's labels are "just nonsensical," "problematic, offensive, and bizarre," examples of "how things can go wrong" with image classification technologies[165].

There are many things with Crawford and Paglen's critique of ImageNet that I agree with. However, I can't help but wonder if their reliance on social constructivism and the subsequent treatment of labels and images as (mis)representations are limiting their criticism. To show this, I will put Crawford and Paglen in conversation with Denise Ferreira da Silva's work on objectivity, subjectivity, and racialized labels[166] and Toril Moi's work on gender and the body[167]. Like Crawford and Paglen, both da Silva and Moi oppose the essentialism of labels,

[160] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019.
[161] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019.
[162] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019
[163] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019
[164] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019
[165] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019
[166]
[167]

especially when applied to human beings. However, unlike Crawford and Paglen, they don't think that essentialism should be opposed with the claim that labels are "socially constructed." I believe we can extend da Silva's and Moi's work on racial and gender classifications to classification technologies like ImageNet.

In *Toward a Global Idea of Race,* da Silva argues against the "quick dismissals of the racial as a scientific concept" and shows why race should not be thought of as a socially constructed concept[168]. Da Silva understands the desire to reject biological explanations of race given that there are no meaningful scientific differences between races, and also that there is a clear track record of how racial differences are "constructed" by nations, corporations, and so forth. The problem, however, is that "cultural difference" gets used in the same way as "genetic difference," consequently essentializing racialized cultures and identities as immutable traits[169]. Da Silva thinks this happens, in part, because of post-modern scholars underestimate the importance of science and the "biological" body to people's understanding of race. She thinks scholars try to escape the "exteriority" of the racialized body by describing the subject in terms of "interiority and temporality."[170] In what she calls the "transparency thesis," da Silva argues that an aversion to "science" and "exteriority" stems from a desire for a transcendental and "transparent" subject - one with no body, no race, and no labels[171].

In many ways, da Silva's "transparency thesis" aligns with Toril Moi's Wittgensteinian-inspired work in *Sex, Gender, and the Body*[172]. Like da Silva, Moi describes the desire among

[168] Da Silva, Denise Ferreira. "Introduction." *Towards a Global Idea of Race.* University of Minnesota Press, 2007, xviiii
[169] Da Silva, Denise Ferreira. "Introduction." *Towards a Global Idea of Race.* University of Minnesota Press, 2007, xxi
[170] Da Silva, Denise Ferreira. "Introduction." *Towards a Global Idea of Race.* University of Minnesota Press, 2007, xxxviii
[171] Da Silva, Denise Ferreira. "Introduction." *Towards a Global Idea of Race.* University of Minnesota Press, 2007, xxxix
[172] Moi, Toril. *Sex, Gender, and the Body.* Oxford University Press, 2005.

some feminists to abandon the concept of "sex" because it is a biologically determined and exclusionary category[173]. Moi points out that the alternative concept of socially constructed gender only tends to replicate the same problems of biological sex, merely substituting the essentialism of biology for that of culture and society[174]. Like da Silva, Moi argues that the desire to reject the biologically sexed body and the exclusive labels it comes with is ultimately a desire for a world with no gender, no language, and no meaning.

In their own ways, da Silva and Moi reject an understanding of gender and race as "socially constructed" concepts. They both pay both special attention to what da Silva calls the "exteriority and spatiality" of the human subject, or the ways in which our physical and material bodies are the conditions of possibility for meaning. Drawing from Moi and da Silva, I believe we can find similar problems with Crawford and Paglen's understanding of labeled images as "socially constructed" objects.

For example, to claim that ImageNet is *not* objective or essentialist, Crawford and Paglen rely on the assumption that one's culture and self-identity are "objective" ways to define the subject. This assumption is a major reason why ImageNet Roulette, their online art piece, works: we are supposed to be simultaneously bothered and reassured by the notion that the offensive labels we receive are not "objectively and scientifically classifying" us, and instead, that subjective labels "are the rule rather than the exception."[175] That is, instead of technology having the power to label us in "objective" ways, Crawford and Paglen try to claim that that *we* do.

The problem with this, as philosopher Roman Amaro points out, is that when we try to "correct" the incoherence between the "artificial" label and our "real" identity, we unwittingly

[173] Moi, Toril. *Sex, Gender, and the Body.* Oxford University Press, 2005, pp 30-31.
[174] Moi, Toril. *Sex, Gender, and the Body.* Oxford University Press, 2005, pp 32.
[175] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019

place “an undue weight on the perception of oneself and environment.”[176] Or as da Silva might say, we end up constructing ourselves in terms of interiority and spatiality rather than exteriority. This is obvious in my own reaction to ImageNet - I opposed the “subjective” “Zionist” label with the “objective” assertion that I am not, *in fact*, a Zionist. In trying to oppose the “false labels” of technology with my “true inner self,” I only ended up essentializing my self-perception as the objective truth.

Thus, Crawford and Paglen don’t escape the clutches of technological essentialism so much as try to shift its burden to other realms, such as self-identity. In this way, their critique seems to reinstate a desire for a “transparent” subject who has no “exteriority” and is only defined by self-perception and “interiority,” as if simply asserting one’s inner identity would change the fact that our exterior selves are constantly being surveilled, classified, and (mis)identified by technology.

The other problem with critiques of science and technology which appeal to “objectivity” is that this reinforces “objectivity” as the only legitimate criteria for truth. And because science can make claims to objectivity in ways that humans – on account of our “subjectivity” – cannot, science has become “the proper domain for the production of the truth of man,” as da Silva puts it[177]. Thus, opposing the science’s objectivity with cultural or personal subjectivity is only “repeating the exclusionary effects of the modern grand narratives of science and history.”[178] For example, the fact that Crawford and Paglen critique ImageNet for being "biased,"

[176] Amaro, Ramon. “As If.” *E-Flux,* 2019. Accessed from https://www.e-flux.com/architecture/becoming-digital/248073/as-if/

[177] Da Silva, Denise Ferreira. “Introduction.” *Towards a Global Idea of Race.* University of Minnesota Press, 2007, xviii

[178] Da Silva, Denise Ferreira. “The Transparency Thesis.” *Towards a Global Idea of Race.* University of Minnesota Press, 2007, xviii

"misrepresentative," or "subjective" seems to imply that ImageNet is kind of object which *could* be unbiased, representative, or objective.

I don't mean to suggest that criticizing ImageNet for being "subjective" *necessarily* implies that it *is* objective; nor do I wish to say calling technology "subjective" is *wrong.* The Wittgensteinian insight here is that our critiques of something can reinforce its operative assumptions. As Wittgenstein puts this idea: "One thinks only of the normal ways in which a mechanism goes wrong, not, say, of cog-wheels suddenly going soft, or penetrating each other, and so on" (§613)[179]. Thus, I want to explore how the logic of the "subjective/objective" binary stems from Crawford and Paglen's picture of technology, namely of what can and cannot go wrong with it.

For example, consider how Crawford and Paglen criticize the "JAFFE" database, a collection of images which classifies facial expressions. They take issue with how the labeled images claim to "depict a woman with an 'angry' facial expression," rather than "the fact" that the woman is "mimicking an angry expression."[180] By critiquing classification technologies for how they "mislabel" images, this reveals an assumption that labels and images are supposed to be, in fact, accurate representations of real life. When Crawford and Paglen claim that ImageNet is "inaccurate," they are reinforcing the assumption that ImageNet is a model "to which reality must conform," in Wittgenstein's words (§613)[181]. The problem with this, as Wittgenstein argues, is that thinking of models only in terms of representation will lead one to make empty assertions which conceal what other work models do in the world. In §613, Wittgenstein writes:

[179] Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe, P. M. S. Hacker, Joachim Schulte. Wiley-Blackwell. 2009, §613
[180] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019
[181] Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe, P. M. S. Hacker, Joachim Schulte. Wiley-Blackwell. 2009. §131

> For we can avoid unfairness or vacuity in our assertions only by presenting the model as what it is, as an object of comparison—as a sort of yard-stick; not as a preconception to which reality *must* conform. (The dogmatism into which we fall so easily in doing philosophy.)[182]

The "dogmatism" of models as representations makes it harder to see ImageNet for everything else it is – for example, an illicit invasion of privacy, a means of reifying racist and sexist institutions, or even material for an artistic piece like ImageNet Roulette. While Crawford and Paglen do touch on some problems with ImageNet outside of a label's accuracy itself – they mention the privacy violation of the original image search and collection process; the low wages and exploitation of Amazon Mechanical Turk workers who label these images; and the weaponization of racist technologies in police departments[183] - they do not adequately show why focusing on representation or accuracy will address any of these material issues.

I find that their critique leaves us at an impasse which is not uncommon in even highly nuanced critiques of classification technologies: the ubiquity of ImageNet in image recognition software means that better labels are desperately needed, while at the same time, making these labels or images more "accurate" or more "objective" would likely require more data collection and increased surveillance of marginalized communities. (For example: in an attempt to make their facial recognition algorithm perform better on dark-skinned faces, Google decided to take photos of homeless Black people living the streets, all under the guise of making their datasets more "fair" and "accurate."[184]) In other words, we want better labels and we also want no labels; but better labels are undesirable because they authorize more surveillance, and no labels aren't possible because ImageNet's classifications have already been absorbed into the world.

[182] Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe, P. M. S. Hacker, Joachim Schulte. Wiley-Blackwell. 2009. §131
[183] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019.
[184] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019.

Holding these two goals in tandem creates a tension which is described at length by Ramon Amaro, who sees a similar phenomenon in another art project, a facial recognition mirror which failed to recognize the face of its Black creator[185]. When the creator of the so-called "Aspire Mirror" called for more inclusive training data, to "code with a universal gaze," Amaro criticized this "necessity to reconcile the psychic potential of the racialized individual with that of a predetermined technical structure."[186] Consequently, Amaro argues that the "black technical object" is too often caught in a "recurrent dialectic," or the oscillation between wanting a better classification scheme and wanting to disrupt the power of classification altogether[187]. While Crawford and Paglen acknowledge this impasse, their own picture of technology might preclude opportunities for getting *out* of it.

Crawford and Paglen close their essay with the remark that classification technologies like ImageNet have the power to "shape the world in their own image."[188] They argue that, to fight against this power, we need to shape the world in our image. But Moi and da Silva tell us that an emphasis on self-identity -- while important – may not actually empower us to oppose the material consequences of labels. If anything, it reflects a desire for a "transparent" and internally determined subject - one who's exteriority cannot be defined by others. I can't help but wonder if this desire for transparency and interiority also applies to our technology. That is, once we remove the "subjective" labels in ImageNet and once we stop applying it to humans, do we hope to have a neutral and transparent technology? Do we think we will be left with no technology? Do we have a desire for technology which somehow does not shape or even reflect the world? I

[185] Amaro, Ramon. "As If." *E-Flux,* 2019. Accessed from https://www.e-flux.com/architecture/becoming-digital/248073/as-if/
[186] Amaro, Ramon. "As If." *E-Flux,* 2019. Accessed from https://www.e-flux.com/architecture/becoming-digital/248073/as-if/
[187] Amaro, Ramon. "As If." *E-Flux,* 2019. Accessed from https://www.e-flux.com/architecture/becoming-digital/248073/as-if/
[188] Crawford and Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." 2019.

explore these questions further in the next example, looking at this picture of a neutral and empty technology, as well as the desire for no technology.

Example 2: Bad Data in, Bad Data out

> ...Anything -- and nothing -- is right." --- And this is the position in which, for example, someone finds himself in ethics or aesthetics when he looks for definitions that correspond to our concepts."
>
> In this sort of predicament: always ask yourself: How did we learn the meaning of this word ("good", for instance?) From what sort of examples? In what language games? Then it will be easier to see that the word must have a family of meanings[189].
>
> (Wittgenstein, *Philosophical Investigations,* §77.)

In everyday conversations about the ethics of technology, I hear a claim that goes something like this: there is no inherently bad technology, just bad uses of technology. Or similarly: bad data in, bad data out. Good data in, good data out. With claims like these, it can feel like the ethics of technology is a completely relative affair. However, Wittgenstein tells us that moral relativity, the confused feeling that "Anything -- and nothing -- is right," (§77) can be clarified by looking at our everyday uses of language. In this example, I do just that. I describe a Twitter conversation about technology ethics which seems to be plagued by moral relativity. However, by looking at how words like "good" and "bad" are used, I show that claims like "biased data in, biased data out" are not so much about relative moral standards, but about avoiding blame and guaranteeing the continued use of technology. Thus, I use this example to bring out a picture of "neutral" technology which is incredibly resistant to charges of unethical or racist behavior, raising questions about the relationship between structures of technology, structural racism, and forms of life.

[189] Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe, P. M. S. Hacker, Joachim Schulte. Wiley-Blackwell. 2009. §77.

For purposes of clarity and privacy, my summary of this Twitter conversation has been edited, shortened, and uses pseudo-names. “AI Guy” represents the voice of a self-described “Artificial General Intelligence Engineer.” “STS scholar” stands for the perspective of several scholars who study Science, Technology, and Society. Their collective Twitter exchange begins with news of a neural network algorithm which labeled an image of a Black man holding a thermometer with “gun” and “firearm,” while it labeled an Asian man holding the same object in the same pose with “electronic device,” “mobile phone,” and “technology.” Alluding to their work on racism in machine learning, one STS scholar noted that it was “infuriating, but not surprising.”

At this point, AI Guy responds to the STS scholar and in several tweets, argues that the neural network itself “is not biased” and “not racist.” To prove this, AI Guy makes a distinction between the “neutral” algorithm, the “bad” input data, and the “biased” output results. He describes the situation in this way: “The result is biased but the code to generate it isn’t. The same code could produce unbiased results if fed with better data. My data cruncher is fine - neutral. Bad data in, bad data out. Good data in, good data out.” When asked why the algorithm itself couldn’t be biased, AI Guy responded: “The neural network performs statistical analysis on whatever is shown to it. It doesn’t care. The same neural network can output biased or unbiased results depending what is shown to it during training.” Therefore, it’s not the algorithm’s fault for producing racist results, “it’s the person feeding the data or the data that’s the problem.” Thus, on AI Guy’s view, this racist algorithm should be understood as “99% a case of bad *use* of technology, not bad technology.”

In response to AI Guy’s claims, STS scholars agreed that “there is no inherently ‘good’ or ‘bad’ data.” At the same time, however, several STS scholars argued that the algorithm itself

should be considered “biased.” One STS scholar noted that the algorithm cannot be considered independent of the data it is applied to; therefore, biased data implies a biased algorithm. Furthermore, for algorithms to work, they have to have “bias,” i.e. treat different kinds of data differently. As one STS scholar wrote: “If that wasn’t the case then they wouldn’t function at all. The point of data science is to get the ‘right’ kind of bias.” However, there is no way for an algorithm to automate only the “right” kinds of helpful bias. As one STS scholar put it: “automation reflects the good, the bad, and the ugly; and to think you can only automate the good is a delusion.” Another STS scholar pointed out the human judgement involved in designing the algorithm itself. As they wrote, “Algorithms are opinions embedded in code. Opinions are biased and therefore algorithms are.” However, just because the algorithm is biased doesn’t mean the algorithm itself is to blame for the racism. One STS scholar agreed with AI Guy that there is “no sense talking of the bias being the algorithms ‘fault’ - it is always people at the end of the day.”

This Twitter conversation is an example of the sorts of exchanges which, in my opinion, are quite common in casual discussions about technology ethics. It begins with a specific case of a racist or sexist algorithm and devolves into a question of who or what is to blame for this so-called “bias.” There are three “language games” of sorts I want to bring out in this example: the separation of data from the algorithm; the questions of blame and responsibility; and the distinction between the “use” of technology and technology itself. After reviewing these various language games, I consider whether they can be situated in a form of life in which racism is a type of “bias” rather than a structural matter.

Let’s start with the differentiation between the data and the code / algorithm. On AI Guy’s view, algorithms are a kind of abstract theory or method, while data is the real-world stuff

it gets applied to. This is why a neural network algorithm is unbiased until the moment it is applied to biased data. At that point, the person who inputs the biased data is to blame for any problematic results - not the algorithm itself. Presumably, had that person put in “good data,” the neural network wouldn’t have produced such a racist result: good data in, good data out. However, this raised the question: can one really separate the code from the data it is applied to? As one STS scholar put it: “Emphasizing the idea that the "core" technical part (e.g. code but not data) is not the problem suggests they're separate. But are they? How do you fix one without fixing the other?” I completely agree with this STS perspective. AI Guy is talking about neural networks like abstract objects, but we know that these kinds of advanced deep learning techniques would not work without lots of data to train them on. Thus, AI Guy is implying that algorithms are a purely technical matter while data is the stuff of humans, society, and ethics. Of course, this separation between the technical / theoretical and the ethical / real world is nothing new. But why make this distinction in this particular case? What are the stakes of AI Guy’s disagreement with the STS scholars?

AI Guy claims he makes the distinction between biased data and neutral algorithms for “*defending* the neural network as a technology that isn’t made with bad biases (or any biases).” This defensive posturing is perhaps the most consistent response I notice from computer scientists, engineers, and AI enthusiasts when I raise similar objections as the STS scholars. While Twitter may not clearly capture the emotions raised in these disagreements, I find that the emotional responses of in-person conversations about racist or biased algorithms are intimately intertwined with feelings of blame and moral responsibility. This suggests that the disagreement over whether algorithms themselves are biased is not just a difference in opinions, as

Wittgenstein might put it, but a disagreement in forms of life (§241)[190]. Think of this way: when someone tells you a joke and you find it offensive, you aren't just simply disagreeing with the logic of the joke. Instead, you get the sense that you and your interlocutor have different ways of living and making judgements, as Toril Moi explains it[191]. In this case, I think the disagreement over where bias in algorithms comes from is less about the nature of algorithms themselves, but about avoiding blame for producing discriminatory technologies. When AI Guy says the neural network itself is neutral, this seems to be a way of "defending" those who build these algorithms (i.e. *him*) from being called racist. On the other hand, when I claim that the algorithm is racist in and of itself, it seems like I am attacking those who build their livelihoods on making these algorithms. This suggests that the seeming relativity of "good data in, good data out" is not so much a true or false claim about the relationship between data and algorithms, but part of language-games such as pointing fingers, shifting blame, and trying to locate the "source" of the problem.

These language games can also tell us something about the concept of technology itself. The defensiveness in these conversations reveals a picture of technology in which technology is never assigned blame. As one STS scholar described this tendency, "It's *never* the technology's fault," so much so that "if something is a problem, by definition it must not be an aspect of the technology." In other words, the neutrality of technology depends on the assumption that problematic data, poor training methods, and bad uses of technology do not count as technical issues. But if technology isn't the data, the training, or the use, what is it? Social and ethical

190 Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe, P. M. S. Hacker, Joachim Schulte. Wiley-Blackwell. 2009. §241.
191 Moi, Toril. *Sex, Gender, and the Body.* Oxford University Press, 2005, p. 4.

problems are already excluded from technology, but this shows that technology itself can be quickly and easily re-constructed to escape social and ethical claims.

The distinction between the technology itself and the *use* of technology also plays an important role in avoiding blame, and more specifically, in ensuring the continued use of technology. When I asked AI Guy to tell me why the difference between "bad use of technology" and "bad technology" mattered, he told me: "'Bad use of technology' implies that we can change how we use technology to get a better outcome. 'Bad technology' is just that and needs changing. It's easier to use Neural Networks better than to redesign Neural Networks and start again." Translation: Bad technology can be fixed with good technology – so while technology can't be blamed, it can still be rewarded. Furthermore, the option of having "no technology" is not really considered. As I told AI Guy, my problem with the "use of technology" is that it implies we *should* use technology to begin with. Perhaps if we could locate the problem in the technology itself, it might be easier to abandon or otherwise abolish harmful technologies.

It's important to remember that this is not just a conversation about technology, but about racism; and how one characterizes racist technologies is also a reflection of how they understand both racism and technology. In this conversation, AI Guy characterizes the racism in the neural network as "bias," and more specifically, *intended* racial bias in the data. But since machines cannot have "intent," this definition makes it impossible to call the algorithm itself "racist." However, this definition of racism clearly ignores structural racism, or racism which is embedded in institutions, laws, and structures. If we think of technology as a kind of structure, then a lack of intent is not a good excuse for racist technologies because they are subject to structural racism, as Ruha Benjamin argues in *Race After Technology: The New Jim Code*[192]. On

[192] Benjamin, Ruha. *Race After Technology: The New Jim Code.* Polity Press, 2019, p. 285.

the other hand, thinking of racism as "interpersonal bias" puts the onus on the individual (or the individuals using the data) to stop being racist.

Interestingly, this *denial* of structural racism can also be found in our language about the use of technology. Thinking of racism as "bias" means that racism can be "fixed" by finding "less biased" data. On the other hand, "fixing" structural racism requires undoing structures and abolishing institutions - including the abolition of technological structures. To prevent racist technologies from simply being abandoned, it makes sense to think of racism as something which comes from *outside* of the technology. Or as AI Guy said, it is easier to "use Neural Networks better" than "redesign Neural Networks and start again." We can link this claim to what Benjamin calls the logic of "new Jim Crow."[193] The "new Jim Crow" refers to the "post-racial" and "color blind" ways in which racial segregation has been maintained, particularly by the extension of slavery into modern prison systems[194]. A defining characteristic of New Jim Crow policies is finding "solutions" or "improvements" for racism which actually maintain the same racist structures. For example, in response to overcrowding in physical jails, digital location trackers and ankle monitoring bracelets have been used to monitor people instead. While this technology might be considered an "improvement," it actually "makes it easier to put people back in; and rather than turn away from the carceral apparatus, it extends it into everyday life."[195] Similarly, thinking of racist technologies as a form of "racial bias" may serve to maintain pre-existing technological structures rather than challenge their underlying logic.

Before I turn to the next example, I want to point to some further areas of investigation between Benjamin and Wittgenstein. The ways in which Benjamin uses "thin description" to talk

[193] Benjamin, Ruha. *Race After Technology: The New Jim Code.* Polity Press, 2019, p. 286.
[194] Benjamin, Ruha. *Race After Technology: The New Jim Code.* Polity Press, 2019, p. 32.
[195] Benjamin, Ruha. *Race After Technology: The New Jim Code.* Polity Press, 2019, p. 286.

about the "reasonableness" of racism or the "facticity of surveillance in Black life"[196] remind me of Wittgenstein's focus on description, agreement in judgements, and forms of life. It would be interesting, then, to think of racism as a form of life. By thinking of racism as a form of life, it isn't just "excluded" from conversations about technology, but the reason why these language games *make sense*. For example, the belief that racism is "bias" could help explain why some don't see racist technology as a structural issue of the algorithm itself. Likewise, the relativity of "bad data in, good bad data out" may also be another way to maintain racist structures and avoid personal blame for racism. Furthermore, the presumed "neutrality" of technology might rely on a form of life related to "color blind" ideology that maintains the new *Jim Code* era. To think of anti-Blackness as the conditions of possibility for both our technology and our ethics would be a fascinating subject to take up. There are all kinds of questions which could be asked. Is the invisibility of technologies such as Amazon Alexa or Roomba Robot related to the invisibility of Black labor? Do we fail to understand and care about digital privacy because the history of surveillance is rooted in tracking Black slaves, Black prisoners, and Black communities? In which ways do we fail to understand technology ethics because we don't locate its history in anti-Black racism? What strategies of resistance do we miss out on because we don't support anti-racist activism? Putting Wittgensteinian in conversation with scholars who link technologies of surveillance and classification to anti-Blackness could provide interesting and clear ways of answering these questions.

Example 3: AI Ethics and Everyday AI

In this section, I look at ordinary and everyday uses of "AI" and AI technologies. Because even though the ability to speak and write has long been criteria for debating the "intelligence" of

196 Benjamin, Ruha. *Race After Technology: The New Jim Code.* Polity Press, 2019, pp. 75-77.

machines, these conversations often do not engage with how actual AI and NLP technologies function. As a result, it is nearly impossible to avoid falling into theoretical debates about whether a machine can think, or whether we will have a future in which robots have rights. In order to compare the reality of AI with how popular culture and philosophers treat it, I turn to a paper by Abeba Birhane and Jelle van Dijk called "Robot Rights? Let's Talk about Human Welfare Instead." In many ways, Birhane and Van Dijk's paper is in the spirit of Wittgenstein. Birhane and Van Dijk begin their paper by shedding light on the assumptions about technology and the human which cause confusions about robot rights and AI ethics. By looking at how both philosophers and ordinary people use the word "AI," they embody Wittgenstein's conviction that "grammar tells us what kind of object anything is" (§373)[197]. Their respect for the everyday and ordinary aspects of life is reflected in their careful attention to "what does exist," namely "machines with software that we call 'AI.'"[198] And furthermore, their critique of the neo-Cartesian viewpoint which underlies AI reminds me of Stanley Cavell's work on Descartes, skepticism, and the "problem of other minds." Nonetheless, I think we can use Wittgenstein and Cavell to push Birhane and Van Dijk's paper even further. In this section, I raise questions about the following things: the philosophical discourse around AI; what it means to pay attention to "everyday" realities of AI; the anthropomorphizing of technology; and the relationship between slavery and machines.

Birhane and Van Dijk begin their paper by looking at the project of "artificial general intelligence" (AGI), the idea that machines can have the full range of human-like intelligence.

[197] Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe, P. M. S. Hacker, Joachim Schulte. Wiley-Blackwell, 2009, §373

[198] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 6. Accessed from https://arxiv.org/pdf/2001.05046.pdf

According to Birhane and Van Dijk, an underlying assumption of AGI is that the brain is a computer. They claim this assumption is based on a "neo-Cartesian" understanding of the self[199]. Unlike traditional Cartesian dualism, this neo-Cartesian view holds that the mind can be mapped on to physical states in the brain. On this view, the mind is "software" embedded in the "hardware" of the brain. Thus, achieving AGI is a technical problem of figuring out how to wire a computer like a human brain. Or as Birhane and Van Dijk put it: "a 'human', on this model, can in principle be 'built', because what it takes to be human is ultimately a particular, complex configuration of physical processes."[200]

However, Birhane and Van Dijk argue that the neo-Cartesian viewpoint is wrong and the assumption that brains are like computers is misleading[201]. I want to quickly preface their critique of the neo-Cartesian view with Stanley Cavell's work in *The Claim of Reason.* In "Skepticism and The Problem of Others," Cavell rejects the comparison between machines and humans not so much on the grounds that machines can't have human sentience, but because this would also open the possibility for humans not to have sentience. In Cavell's words:

> We should not have accepted the vision of a machine's having a region of sentience in the first place, because: "If the human body is a machine then a machine has sentience" is no better an inference than "If the human body is a machine then the body does not have sentience."[202]

Cavell argues that building a machine based on the "perfect human body" is based on the assumption that the body is "inhabited" by the mind and therefore can be separated from it. As

199 Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 2. Accessed from https://arxiv.org/pdf/2001.05046.pdf

200 Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 2. Accessed from https://arxiv.org/pdf/2001.05046.pdf

201 Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 3. Accessed from https://arxiv.org/pdf/2001.05046.pdf

202 Cavell, Stanley. "Skepticism and The Problem of Others." *The Claim of Reason.* p. 414

Cavell puts it: "To take the human body as a machine is as much, or little, a vision of humanity as the vision that takes the body to be inhabited by something *else*."[203] Contrary to this view, then, Cavell argues that the human mind is not only inseparable from the human body, but is also connected with others' minds and bodies.

Birhane and Van Dijk also reject the "illusion" that the human mind can be detached from our body and the "natural, artificial and social world that we live in."[204] Instead, they propose an "embodied" and "extended" view of cognition, one in which a human being is "fully embedded and enmeshed with our designed surroundings."[205] In their words:

> We take a post-Cartesian, phenomenological view in which being human means having a lived embodied experience, which itself is embedded in social practices. Technological artifacts form a crucial part of this being, yet artifacts themselves are not that same kind of being. The relation between human and technology is tightly intertwined, but not symmetrical.[206]

Birhane and Van Dijk emphasize that the embeddedness of humans with other objects doesn't entail that machines and humans are the same or even similar types of being. Thus, we cannot conflate the human, the brain, and the computer. In fact, Birhane and Van Dijk argue that this equivalence is what leads to serious "problems" in the philosophy of AI.

If brains are computers, then robots could be "very much like ourselves," raising the question of whether robots should have the same rights and privileges that humans do. The problem with this line of thinking, however, is that there currently *are* no robots which actually resemble humans. Thus, "debating the necessary conditions for robot rights keeps putting focus

[203] Cavell, Stanley. "Skepticism and The Problem of Others." *The Claim of Reason.* P. 414
[204] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 3. Accessed from https://arxiv.org/pdf/2001.05046.pdf
[205] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 3. Accessed from https://arxiv.org/pdf/2001.05046.pdf
[206] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 3. Accessed from https://arxiv.org/pdf/2001.05046.pdf

on (non-existent) machines, instead of on real people."[207] By speculating about "future visions of sentient machines," philosophers end up ignoring how these machines currently affect sentient humans[208]. However, Birhane and Van Dijk make it clear that the *stakes* of their disagreement with AI enthusiasts and robot rights proponents is not whether "artificial general intelligence" *will* exist, but how speculation into the matter can distract from the ways in which actual AI systems undermine human welfare.

Accordingly, Birhane and Van Dijk turn their attention to AI technologies which *do* exist, showing how these technologies impact human lives. They describe how the profit-driven sectors who build and deploy AI systems treat moral and political issues like technical problems which can be quantified and automated. They touch on racism in predictive policing systems; sexism in the display of STEM career ads; racism in recidivism algorithms; bias in the politics of search engines; and racism in medicine[209]. Birhane and Van Dijk also describe the dehumanizing labor of Amazon Mechanical Turk, in which Amazon pays people to perform thousands of microtasks for low wages[210]. From the "perpetuation of historical and social bias and injustice, to invasion of privacy, to exploitation of human labour," Birhane and Van Dijk clearly demonstrate how AI hurts human welfare[211].

[207] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 4. Accessed from https://arxiv.org/pdf/2001.05046.pdf
[208] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 4. Accessed from https://arxiv.org/pdf/2001.05046.pdf

[209] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 6. Accessed from https://arxiv.org/pdf/2001.05046.pdf

[210] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 6. Accessed from https://arxiv.org/pdf/2001.05046.pdf
[211] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 6. Accessed from https://arxiv.org/pdf/2001.05046.pdf

Birhane and Van Dijk argue that AI is particularly dangerous because it is both ubiquitous and covert. On their view, the most profitable and most dangerous technologies are those which "weave themselves into the fabric of everyday life until they are indistinguishable from it."[212] For example, seemingly ordinary technologies such as Amazon Alexa and the Roomba Vacuum are "designed to render all corners of lived experience as behavioral data."[213] According to Birhane and Van Dijk, AI is made even more invisible by the failure of academics and pop culture alike to accurately portray it. This is why speculation about non-existent AI contributes to a norm of "techno-optimism" which conceals "the current development of dehumanizing technological infrastructure."[214] Therefore, an important step out of the shadows of AI is to use language which accurately describes it. According to Birhane and Van Dijk, it is not "valid" to perceive artifacts either as "mere machines" or "intelligent others," "*even if we socially talk about them that way*."[215] They claim that both of these interpretations underestimate how entrenched we are with our technology and fail to do justice to its ordinary occurrences.

On one hand, I agree with Birhane and Van Dijk that the way we talk about AI and technology can be seriously misleading. I also agree that we underestimate our embeddedness with the technological and social world. On the other hand, I'm not sure if we should dismiss the ways we "socially talk" about AI as "invalid" or "wrong." Drawing from Wittgenstein, it only makes sense to call our language "valid" or "invalid" is we are making a claim to validity (§136).

212 Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 5. Accessed from https://arxiv.org/pdf/2001.05046.pdf

213 Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 5. Accessed from https://arxiv.org/pdf/2001.05046.pdf

214 Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 6. Accessed from https://arxiv.org/pdf/2001.05046.pdf

215 Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 5. Accessed from https://arxiv.org/pdf/2001.05046.pdf

In other words, the "misleading" ways people talk about AI are only "invalid" if those people are, in every instance, putting forth a definition of AI or claiming to accurately describe it – and I'm not sure if this is case with "AI." Wittgenstein emphasizes that we do much more with our words than put forth true or false propositions and even Birhane and Van Dijk themselves point out all the different ways we use "AI." "AI" is used to *distract* from the human labor behind it; to *advertise* technologies; to *overemphasize* the sophistication of machine learning methods. Unlike Birhane and Van Dijk, then, I don't think the problem with "AI" is located so much in our language, but in our world: namely, the unregulated havoc that techno-capitalism wrecks on humans. I believe that even the "wrong" ways we use AI can *accurately* reflect what kind of object and concept it is.

With this in mind, turning back to the question of what counts as a "mere machine" versus an "intelligent being" is not only an inquiry into what actually is a mere machine or an intelligent being, but what work those terms do in which situations. I want to briefly focus on one situation in which the distinction between "mere machine" and "intelligent being" has serious consequences: the dehumanization and treatment of Black people during and after the transatlantic slave trade. The treatment of Black people as slaves, machines, and capital has been articulated and explored by a wide variety of scholarship. I will give a high-level overview of some of these perspectives and use Wittgenstein and Cavell to raise several questions about the distinctions or comparisons we make between machines, humans, and Black slaves. These questions, in turn, can further help us understand the divide between technology and the human.

As a starting point, let's briefly return to what Birhane and Van Dijk have to say about the connection between slaves, machines, and dehumanization. They argue that slaves should *not*

be considered machines because "slaves are *humans* abused as machines."[216] Likewise, machines should *not* be considered slaves because machines are not humans and cannot be dehumanized in the first place. Thus, the comparison between machines and humans only serves to dehumanize slaves. As Birhane and Van Dijk write:

> The robot is the very model against which we judge whether humans are dehumanized... By putting actual slaves, women, and 'other races' in one list with robots, one does not humanize them all, one dehumanizes the actual humans in the list.[217]

According to Birhane and Van Dijk, robots are the "model" against which we construct the category of human. However, others have also argued that *Blackness* is what defines the limits of the "human." I want to explore the idea that the two are not mutually exclusive - that is, Blackness does not simply replace robots as a stand-in for the "inhuman" (or vice-versa,) but that they co-constitute each other.

One way to explore this idea is by engaging with afropessimism. Afropessimism, to put is simply, argues that the condition of Blackness is unique and incomparable to other forms of marginalization because of the history of slavery. While Birhane and Van Dijk include "slaves, women and 'other races'" on the same list as examples of humans who have been treated like machines, afropessimists would argue that Black slaves are not only treated like machines, but mark the boundary between humans and machines. And rather than try to include Black slaves into the category of human, afropessimism begins with the premise that this will never happen. As artist and writer Aria Dean argues, the Black slave will always represent the "inhuman," or

[216] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 3. Accessed from https://arxiv.org/pdf/2001.05046.pdf
[217] Birhane, Abeba and Van Dijk, Jelle. "Robot Rights? Let's Talk about Human Welfare Instead." 2020, p. 3. Accessed from https://arxiv.org/pdf/2001.05046.pdf

what she calls the Black (non)subject[218]. The construction of this "human-but-not," "subject-object," "techno-human," "was-African-made-Slave" comes from the development of racial capitalism from the Middle Passage onwards[219]. Racial capitalism is a historical and theoretical framework developed by Cedric Robinson which does not separate questions of capitalism from questions of race. It is based on the fact that Western capitalism (and thus the development of modern technology) was "kick-started by the rape of the African continent," as theorist Frank Wilderson puts it[220]. According to Wilderson, the very concept of capital is derived from approaching a "was-African-made-Black" body with "direct relations of force," rather than "approaching a white body with variable capital."[221] As a result, the Black slave simultaneously makes up both capital and subjecthood, both machine and human. In other words, Blackness is not simply excluded from the category nor simply comparable to a machine. Instead, Blackness what produces those concepts in the first place.

It is worth quoting Wilderson to further clarify this idea of the Black slave as the Black (non)subject:

> …. as an ontological position, that is, as a grammar of suffering, the Slave is not a laborer but an anti-Human, a positionality against which Humanity establishes, maintains, and renews it coherence, its corporeal integrity… the Slave is, to borrow from Patterson, generally dishonored, perpetually open to gratuitous

218 Dean, Aria. "Notes on Blaccelerationism." *E-flux,* 2017. Accessed from https://www.e-flux.com/journal/87/169402/notes-on-blacceleration/
219 Dean, Aria. "Notes on Blaccelerationism." *E-flux,* 2017. Accessed from https://www.e-flux.com/journal/87/169402/notes-on-blacceleration/
220 Wilderson, Frank. "Gramsci's Black Marx, Whither the Slave in Civil Society?" *Social Identities, Volume 9, Number 2,* 2009, p. 229.
221 Wilderson, Frank. "Gramsci's Black Marx, Whither the Slave in Civil Society?" *Social Identities, Volume 9, Number 2,* 2009, p. 229.

> violence, and void of kinship structure, that is, having no relations that need be recognized, a being outside of relationality[222].

As Wilderson theorizes it, the organizing feature of the Black slave is not forced labor, (since not all slaves had to work, and this characterizes slavery too broadly) but rather the concept of social death[223]. Drawing from the realities of the transatlantic slave trade, Wilderson argues that social death has three elements. The first element is "gratuitous violence," or the idea that the slave's body is structurally and perpetually vulnerable and open to violence. The second element is that the slave is "void of kinship structure...having no relations that need be recognized."[224] The slave's family might exist hypothetically, but these familial relations have not and will not be recognized by society. The third element is the concept of "dishonor," in which Blackness is the a priori state of being dishonored, before any action warrants it[225]. If we are to follow Saidiya Hartman's contention that Blackness remains the "afterlife of slavery," then the three elements of social death which define the Black slave still define Blackness today[226].

While Birhane and Van Dijk argue that slaves *are* humans treated as machines, afro-pessimism argues that Black slaves *were* once humans treated as machines, but that centuries of anti-Black racism has effectively transformed Black people into the realm of machine-human. At first glance, it seems like Wittgenstein and Cavell - who often come off as committed humanists - would hold more in common with Birhane and Van Dijk than the afropessimist perspective. However, this may not be the case. On Wittgenstein's view, the difference between a human and

222 Wilderson, Frank. *Red, White & Black: Cinema and the Structure of US Antagonisms*. Duke University Press, 2010, p. 19.
223 Wilderson, Frank "Blacks and the Master/Slave Condition," (Interview by C.S. Soong, 2015), 18.
224 Wilderson, Frank. *Red, White & Black: Cinema and the Structure of US Antagonisms*. Duke University Press, 2010, p. 19.
225 Frank Wilderson, "Blacks and the Master/Slave Condition," 18.
226 Saidiya Hartman, "The Belly of the World: A Note on Black Women's Labors," *Souls* 18, no. 1, 2016, 83.

a machine can be captured by the way we treat humans as if they have a soul. In the later part of the *Philosophical Investigations,* Wittgenstein writes:

> Suppose I say of a friend: "He isn't an automaton".—What information is conveyed by this, and to whom would it be information? To a human being who meets him in ordinary circumstances? What information could it give him? (At the very most that this man always behaves like a human being, and not occasionally like a machine.) "I believe that he is not an automaton", just like that, so far makes no sense.
>
> My attitude towards him is an attitude towards a soul. I am not of the opinion that he has a soul.[227]

According to Wittgenstein, we do not "hold" a "belief" that one has a soul, but rather *treat* someone or *acknowledge* them as if they do. On the other hand, we do not treat machines as if they have souls. However, as Stanley Cavell and Stephanie Spoto have argued, white people also tend to display a "soul blindness" towards Black people. Investigating how these two ways of "not seeing souls" are related could make use of afropessimism in interesting ways.

In "Wittgenstein, Aspect Blindness, and White Supremacy,"[228] Stephanie Spoto argues that white people have "soul blindness" towards Black people. She develops this idea using Wittgenstein's notion of "aspect blindness." As Wittgenstein writes:

> Could there be human beings lacking the ability to see something as something—what would that be like? What sort of consequences would it have?—Would this defect be comparable to colour-blindness, or to not having absolute pitch?—We will call it "aspect blindness."[229]

[227] Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe, P. M. S. Hacker, Joachim Schulte. Wiley-Blackwell, 2009, §178*.
[228] Spoto, Stephanie. "Wittgenstein, Aspect Blindness, and White Supremacy." Volume 7, Issue 2, 2019, pp. 247-260
[229] Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe, P. M. S. Hacker, Joachim Schulte. Wiley-Blackwell, 2009, §179*.

Aspect blindness, Spoto argues, can help us explore how white people do and do not acknowledge racial differences. For example, the ways in which white people *see* Blackness, but fail to *acknowledge* the souls of Black people. Spoto emphasizes that soul blindness is "not in the biological sense," as Spoto puts it, but "in some kind of moral or ethical sense."[230] But what if it *was* in a biological sense? What if white people treat Black people more like automatons with souls rather than humans who lack them? Or perhaps as something in between – a form of artificial intelligence in its own right? How does the Black (in)human and the technological (in)human come together in our picture of AI? As Louis Chude-Sokei argues, many philosophical and cultural understandings of artificial intelligence have been deeply shaped by Black (in)humanity[231]. Looking at science-fiction novels and the rise of "cybernetics" in late nineteenth-century England and the United States, Chude-Sokei shows that questions of whether a machine can think or have feelings are popularized at times when slave rebellions and abolition movements threatened to humanize Black slaves. We might wonder, then, how our current obsession with AI is a reflection of our contemporary racial landscape. Putting Wittgenstein in conversation with afro-pessimism, then, might be one way to do this.

[230] Spoto, Stephanie. "Wittgenstein, Aspect Blindness, and White Supremacy." Volume 7, Issue 2, 2019, pp. 247-260

[231] Chude-Sokei, L. "Humanizing the Machine" (Chapter 3) The Sound of Culture. Diaspora and Black Technopoietics, 2016.