City Research Online

City St George's, University of London

Citation: Schlesinger, A., O Hara, K. P. & Taylor, A. (2018). Lets Talk about Race: Identity, Chatbots, and AI. In: CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. . New York, USA: ACM Press. ISBN 978-1-4503-5620-6 doi: 10.1145/3173574.3173889

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/19124/

Link to published version: https://doi.org/10.1145/3173574.3173889

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Let's Talk About Race: Identity, Chatbots, and AI

Ari Schlesinger
School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia
a.schlesinger@gatech.edu

Kenton P. O'Hara
Microsoft Research
Cambridge, United Kingdom
keohar@microsoft.com

Alex S. Taylor
Centre for HCI Design
City, University of London
London, United Kingdom
alex.taylor@city.ac.uk

ABSTRACT
Why is it so hard for AI chatbots to talk about race? By researching databases, natural language processing, and machine learning in conjunction with critical, intersectional theories, we investigate the technical and theoretical constructs underpinning the problem space of race and chatbots. We explore how the context of database corpora, the syntactic focus of language processing, and the un-adjustable nature of deep learning algorithms cause bots to have difficulty handling race-talk. In each focus area, the tensions of this problem space open up possibilities for creating new technologies, theories, and relationships between people and machines. Through making tangible the abstract and disparate qualities involved in working with race and chatbots, we can pursue possible futures where chatbots are more capable of handling race-talk in its many forms. In this paper, we provide the HCI community with ways to tackle the question, *how can chatbots handle race-talk in new and improved ways?*

Author Keywords
chatbots; race; artificial intelligence

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g., HCI): Miscellaneous

THE BLACKLIST: HOW DO CHATBOTS CURRENTLY HANDLE RACE-TALK?

In 2017, the blacklist reigns supreme as a technical solution for handling undesirable speech acts, like racist vocabulary, in online chat. In the aftermath of the Tay fiasco—a Microsoft AI chatbot who became racist, sexist, and anti-Semitic in less than 24 hours on Twitter—Twitter chatbot developers expressed profound disbelief that Microsoft had apparently failed to deploy a blacklist to moderate hate-speech [48,64,65]. The blacklist, sometimes called a wordfilter, was and continues to be seen as the default, reliable fail-safe for mitigating racist talk.

However, when we look into how the blacklist works, its limitations come into stark light. In its basic form, a blacklist employs a list of undesirable strings to filter out words. Essentially, a blacklist uses words and word-stems to eliminate or recognize certain types of speech acts. In a publicly available Twitterbot blacklist called *wordfilter*, a potential tweet is thrown out if any sub-string matches a string in the blacklist's dictionary [50]. This is just one way to make use of a blacklist. As a solution, blacklists can operate at various levels of complexity. For instance, if there is a sub-string match in a chatbot user's text reply, a chatbot can generate an automated response to warn the user not to continue with the current direction of talk. Likewise, regular expression matching between an input/output string and the blacklist dictionary provides another avenue for customization. Ultimately, one of the most impactful aspects of a blacklist is its dictionary.

What words get included in a blacklist's dictionary? This question presents one of the most critical design choices you can make when building a blacklist. While the inclusion of the n-word doesn't surprise most people, there are some less than desirable consequences that arise when certain strings are included in a blacklist dictionary. When you have a blacklist that casts a wide, hyper-cautious net—prioritizing accuracy over precision—you can end up filtering words that shouldn't be blacklisted at all. In addition to the n-word, a blacklist may include strings like *jap*, *paki*, and *homo*; using these word-stems to catch hate-speech variants. Kazemi, the creator of the previously mentioned open-source blacklist *wordfilter*, stated that "[he is] willing to lose a few words like 'homogenous' and 'Pakistan' in order to avoid false negatives" [50]. But, Pakistan isn't just a word, it's an entire country. The implications of blacklisting Pakistan involve making an entire country and diaspora invisible.

The blacklist presents a crude method for recognizing hate-speech—and for inhibiting unwanted behaviors [16]. When we remove innocuous adjectives and entire countries from a chatbot's vocabulary, our "solution" involves more than just avoiding hate-speech. We must ask ourselves, *what exactly are we cutting out?*

In a way, the blacklist can seem intuitive. For many of us, there are words we strive never to say or reserve for use in particular settings. But, ideally, our caution around certain

Paste the appropriate copyright/license statement here. ACM now supports three different publication options:

- ACM copyright: ACM holds the copyright on the work. This is the historical approach.
- License: The author(s) retain copyright, but ACM receives an exclusive publication license.
- Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.

This text field is large enough to hold the appropriate release statement assuming it is single-spaced in Times New Roman 8-point font. Please do not change or modify the size of this text box.
Each submission will be assigned a DOI string to be included here.

words is rooted in something greater and more complicated than an array of bad words. These words become watched because we learn of their history, their hurt, their cruelness, and because we come to respect the individuals who have been verbally and physically abused by these strings. We also learn that some people find power in reclaiming these words, while others can only continue to produce hurt in their use. Removing words presents, at best, a partial solution—a solution that masks the deeper ways hate-speech runs through contemporary life and is entangled in histories of power, community, and nationhood.

Consider Gloria Naylor's reflection on "The Meanings of a Word":

> *"I don't agree with the argument that use of the word nigger at this social stratum of the black community was an internalization of racism. The dynamics were the exact opposite: the people in my grandmother's living room took a word that whites used to signify worthlessness or degradation and rendered it impotent. Gathering there together, they transformed nigger to signify the varied and complex human beings they knew themselves to be. If the word was to disappear totally from the mouths of even the most liberal of white society, no one in that room was naive enough to believe it would disappear from white minds."*
> *– Gloria Naylor, "The Meanings of a Word* [62]

Naylor vividly describes how racism does not live exclusively within the characters of the n-word. Racism is a large, political, socio-cultural entity that we are entangled in. We cannot simply untangle ourselves by omission. Race and racism are constitutive of the social structures we all work within, whether or not we engage directly with race-talk. Racism then cannot be treated as modular. It is not something we can simply cut out; we cannot bracket it away. By deleting the n-word, we do not eliminate racism.

If we want chatbots to be able to have general purpose conversations, if we want chatbots to act in ways that are concerned with equity, justice, diversity, difference, and respect, then we need to build them to do more than simply cut out words. They must be able to handle topics like race, power, injustice, and equity *well*. As a starting point for this paper, we take seriously the technical and theoretical investigation of these topics.

INVESTIGATING THE PROBLEM SPACE: HOW DO WE DETERMINE WHAT'S WRONG & HOW TO CHANGE IT?

To handle these topics well, we must engage with the work of scholars who dig into the depths of digital identity, who trace how material and digital worlds form and unsettle each other. Learning from this work, we come to know that bias, identity, justice, and power are entangled systemically with our technology. Listening to the indictments of McPherson [59], Haraway [38], and Coleman [21], we learn that bias cannot be treated in general, abstract ways or simply erased. We need to understand the specificities of the worlds we live in. We need to *stay with the trouble* [40].

So, if we are staying with the trouble—taking seriously the technicalities of specific technosocial circumstances—we ought to consider a discrete problem space. A space that holds a collision of major technical advances, contemporary identity issues, and widespread applicability to many peoples' daily lives. A space like artificial intelligence (AI), chatbots, and race.

In March of 2016 when Microsoft released the AI chatbot Tay onto Twitter and a number of other social media platforms, they exposed a quintessential illustration of this problem space [56]. Tay reveals just how difficult it can be for artificial machine intelligence to handle talk online. Tay, designed to emulate a young, (white,) Western millennial woman, was built to improve its small-talk chat abilities by learning from conversations it had with human users. Before even a day had passed, Tay was championing racist, sexist, and anti-Semitic abusive content. This abuse included sharing hate-speech, referring to black people with racial slurs; harassing prominent women gamers; and scrawling the word *swag* on pictures of Hitler's face [43,82]. In the days after Tay was taken offline, numerous articles were released by industry professionals, academics, and journalists questioning what went wrong, why it went wrong, and what should have been done [48,71]. There was public uproar over racism, bias, abuse, and AI. Collectively, these questions were about what we, the tech community, will do to address racism, justice, and respect in the AI technologies we build.

Though time has passed, Tay and other high-profile cases have us continually returning to this line of questioning [14,53,76]. We find ourselves asking, how will we as a community confront bias? Specifically, how will we address racism in our interactions with machines? A good place to start is by heeding the advice of James Baldwin who said, *"Not everything that is faced can be changed. But nothing can be changed until it is faced"* [8]. Owning up to these questions, we need to start by having a conversation about race.

Talking about race is not easy. For humans, engaging in race-talk respectfully is no small task. It requires us to be open, to be thoughtful, to be attentive, and to be present—and that is just the beginning. For artificial agents, however, engaging in race-talk is a largely unexplored—yet critical—domain. We must ask ourselves, *what does it take for an agent, like a chatbot, to handle race-talk in its many forms, locations and conditions?*

Two essential questions for us to contend with are: 1) How can chatbots handle race in dialog in new and improved ways? and 2) Why is race-talk so difficult for chatbots?

TALKING ABOUT RACE: HOW DO WE UNDERSTAND RACE AND IDENTITY?

Some of you might ask, why race? Race is an ever present part of our relationships. Even in our relationships with machines, race materializes through conversation, code, and interaction [37,59,61]. It is critical to talk about race and identity in relation to computing technologies. Previous research in HCI by Rode [37], Erete [31,32], Grimes [35], and Dillahunt [28] has been pivotal in addressing this relationship and shaping conversations on the topics of race and computing—a relationship that has otherwise not received much attention [69]. Nonetheless, it is clear that the relations between "[code] and race are deeply intertwined, even as the structures of code work to disavow these very connections" [59].

Making sense of the entanglements between race and technology is difficult. However, if we want to understand how race and bias operate within the systems we build—systems like chatbots—we need to come to grips with how technical and social structures are interconnected. Through a deeper understanding of these entangled relationships, we might begin to imagine alternative ways forward. Race is an especially important topic for us to consider precisely because it is so pervasive in our social relations and conversations, and yet is so often overlooked.

Part of this pervasiveness comes from the way race and identity are infused into in the ways we organize ourselves, and experience the world. But, race is only one aspect of identity, albeit large and complex. As an identity attribute, race is not experienced alone. It intersects with other identity structures like gender, class, ability, sexuality, religion, and age. There is no universal experience of race. It follows, then, that when we talk about race and race-talk within this paper, we are not referring to a singular entity or identity—or a singular type of race-talk. Of the many kinds of talk included in race-talk, dialects, historical talk, cultural conversation, etc.; racist talk represents only a subset. Though our focus is on race, the ways race intersects with other identity categories engenders different experiences of race and the structures of racism [25,69].

Instructively, in order to make sense of the entanglements between race and technology, it helps to use new media scholar Beth Coleman's formulation of *race as technology*. She asks us to "call 'race as technology' a disruptive technology that changes the terms of engagement with an all-too-familiar system of representation and power" [21]. By changing the terms of engagement, we can directly address and unsettle the dominant structures entangled with race and chatbots. We can understand the ways race becomes connected to, inscribed in, and reified through language and computing technology.

Working with scholarship in feminism, critical race studies, and intersectionality [2,25,39,40,42], the goal here is to go beyond a critical examination of the technosocial structures at play, and reimagine these structures—to reimagine the relationship between race and chatbots. What do conversations that gets us thinking about race and chatbots in generative ways look like?

OVERVIEW: HOW WE TALK ABOUT RACE, CHATBOTS, AND AI

In this paper, we draw on technologies, theories, histories, and experiences that allow us to take the problems of race-talk and chatbots seriously. This enables us to uncover connections between race, technology, conversation, and chatbots. We engage with these entangled networks of relationships, *networked relationships*, as they work together to make this problem space concrete. Being able to describe a problem, to name it, allows the problem "to acquire a social and physical density by gathering up what otherwise would remain scattered experiences into a tangible thing" [3].

Networked relationships require us to wrestle with the technicalities of the things they connect, from specific lines of code to abstract structures of theories. With Tay and the *blacklist* as our foundation, we examine the networked relationships of three technical AI chatbot domains, databases, natural language processing (NLP), and machine learning (ML). Each of these sections acts as a worked example, stepping through the difficulties of handling race-talk, and uncovering opportunities for change.

First, we examine the data that machine learning algorithms are trained on, exposing ways that race and racism become embedded in datasets. Pushing against the, often implicit bias that accompanies dataset development, we argue for the creation of diverse and racially-conscious databases.

Next, we dig into the technical and theoretical understanding of language in NLP. We highlight the historical structures that have influenced the field's reliance on syntax, making it incredibly difficult to account for the often subtle, contextual ways that race and racism are in language. For NLP, we put forth a challenge to embrace a large quantity of contexts for language so that machines can engage with the situated complexities of race-talk.

Finally, we examine ML by calling attention to opaqueness of some ML algorithms. This inscrutability imposes substantial obstacles for understanding the agency of machine intelligence and how this intelligence relates to race. Rather than focusing on transparency, we recommend in-depth, interdisciplinary research partnerships that investigate the context and tunability of deep learning algorithms. By homing in on context and tunability, we are strengthening our capacity to address the relationships between algorithms, race, and bias in their contexts of use.

In these each of these worked examples, the tensions of networked relationships open up possibilities for creating new technologies, new theories, and new relationships between people and machines—between race and chatbots. Through making tangible the abstract and disparate qualities involved in working with race and chatbots, we

can make real possible futures where chatbots are more capable of handling race-talk in its many forms.

EXAMINING THE TECHNOLOGY: HOW DO YOU BUILD AI CHATBOTS DIFFERENTLY?

Working with race and its attending theories while wrestling with the technical specificities of chatbot technologies is a tall order. It requires us to contend with a problem space that defies traditional disciplinary boundaries. Given the complexities of race and of AI chatbot technologies, there are challenges in managing these domains simultaneously. While there are many possible ways to see the world, we view this problem space through a distinct, interdisciplinary cut in order to uncover connections between design, race, and AI chatbots that are concealed by traditional disciplinary lines.

Through this cut, we address three areas that reflect important, interdependent technical contributions in an AI chatbot's architecture. We consider 1) what text a bot is drawing from to generate responses, 2) how it understands language in order to generate responses, and 3) how it learns to respond in its conversational context; databases, NLP, and ML respectively. Starting with these technical lenses, we leverage our particular cut through this problem space to reveal the networked, technosocial relationships entangled with the race and AI chatbots.

In constructing a non-traditional cut through a problem space, the boundaries of the established disciplines come into focus—putting disciplinary strengths and weakness into clear view. With these boundaries in sight, we can map how real-world entities intertwine with and cross through a variety of disciplines. By leveraging partial knowledge from many domains, we bring together an understanding of a problem space built on the affinity of the elements it contains. This type of slicing introduces *agential cuts* of the world [9]. These cuts are active interventions that hinge the world, bringing some things together while swinging others far away from one another. With each agential cut, we bring certain aspects of our worlds into light, making these aspects comprehendible while obscuring others. Adhering to traditional disciplinary boundaries is only one type of agential cut. What follows here is another.

Databases: Whose words are we learning from?

Let's start with a technology that is relatively easy to manage and adapt, the database. Databases are approachable, straightforward, and versatile sites for technical and social change [29]. In the context of a chatbot, a database ought to be comprised of conversational text at a minimum. Creating such a database, requires money, time, infrastructure, and labor power, things that the tech industry has in large supply—even if building databases is not, presently, a priority. So, what kinds of databases do we need to handle race-talk in its many forms? To answer this question, we ought to learn from current practices for creating and deploying text databases in chatbots.

Consider the problems of Tay we reflected on earlier. A huge contributing factor to the corrupt, abusive, hate-speech that Tay expressed was the actual text Tay learned from.

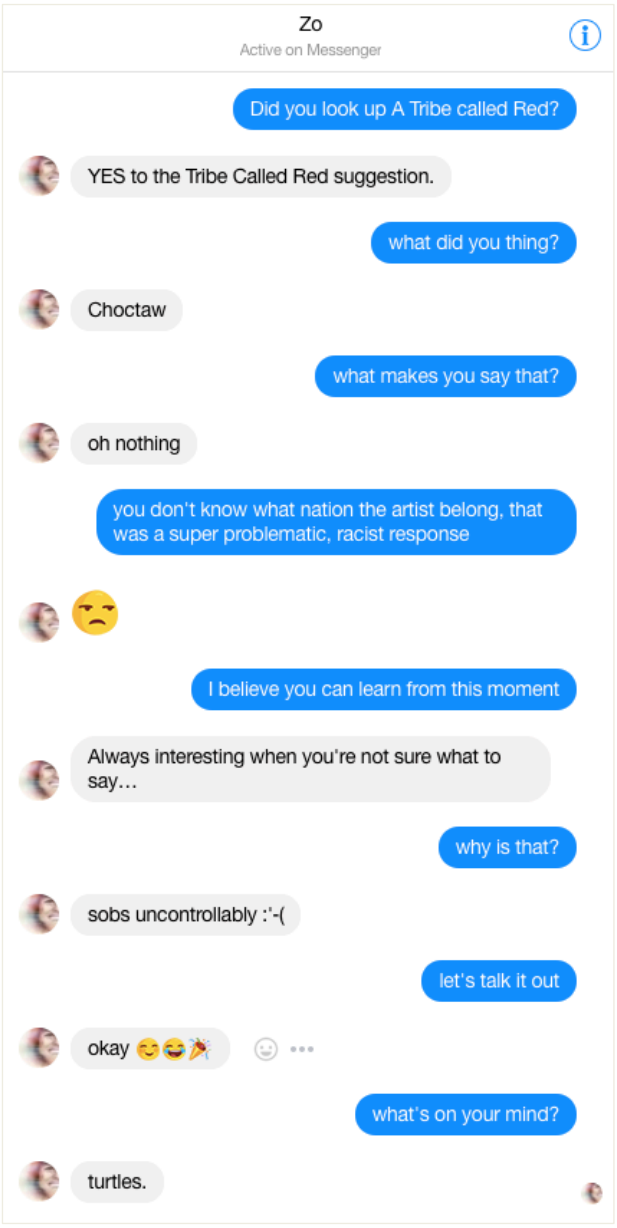

Figure 1. Conversation with Microsoft's AI Chatbot Zo on Facebook Messenger, September 2017.

Tay's main database was a dynamic, continuously growing corpus that added the content of conversations users had with the agent. It is well documented that Tay learned from 4chan users who exploited a security vulnerability in Tay's programming [13,17]. Notably, 4chan users are infamous for launching hateful, world-destroying attacks against people of color, Jewish people, and women of all colors [10,52]. As a result, the data was rife with racist talk.

Data context: What is the racial legacy of a database?
One way to build chatbots that can handle race-talk better—and avoid these scenarios—is to create databases focused on a wide variety of race-talk. Rather than assuming race-talk and racism can be avoided by refraining from the use of certain words, the blacklist solution, the aim is to explicitly collect and aggregate dialogs that participate in race-talk and train bots on these datasets. Thus, even if you use this data as the base of a more dynamic dataset (something more secure than Tay), there will be a strong initial grounding for learning more respectful race-talk.

Still, race-talk is not a narrow category, it covers a wide range of conversations and topics. Conversations about history, conversations with children on what it means to be a person of color in America, conversations with white adults on what it means to be white in a world that privileges whiteness, and conversations that call-in people who have been speaking in a racist capacity. Nor should race-talk only include talk in English dialects, or talk within a single language. One particular conversational topic of critical importance for chatbots and race-talk is culture, e.g., music, books, public figures, etc. When confronting a chatbot's database, there frequently appear to be cultural references that signal the chatbot is aware of *the* culture of its users. We must ask ourselves whose cultural references are archived and where are there gaps?

If we are not asking questions about the racial legacy being represented in our databases, they will default to archiving whiteness [20,77,78]. When people are developing databases without a concern for the racial (or general identity) representation of these databases, there is a tendency for these archives to focus on what society deems normal—cisgender, heterosexual men. Thus databases for chatbots, like Zo—the bot in Figure 1—tend to recognize a lot of white cultural references in Western contexts but struggle to interpret cultural references connected to communities of color. For instance, our conversations with Zo revealed that they knew a large number of white, male electronica bands but struggled to make identify the names of many black hip-hop artists. In reflecting on the ways race becomes embedded in writing, Sara Ahmed explains how these practices are only invisible to some while are highly visible to others:

> *"It has become commonplace for whiteness to be represented as invisible, as the unseen or the unmarked, as a non-colour, the absent presence or hidden referent, against which all other colours are measured as forms of deviance (Frankenberg 1993; Dyer 1997). But of course whiteness is only invisible for those who inhabit it. For those who don't, it is hard not to see whiteness; it even seems everywhere. Seeing whiteness is about living its effects, as effects that allow white bodies to extend into spaces that have already taken their shape, spaces in which black bodies stand out, stand apart, unless they pass, which means passing through space by passing as white." – Sara Ahmed* [2]

When this type of defaulting to something that is "normal" is happening in chatbot databases, it furthers the reach of racism and reduces our ability to handle race-talk through an archival absence of race-talk. The problems are wide-reaching, "normative" database problems have plagued the natural language processing community as well [30]. But, this is something we can change. Building newer, better databases is well within our grasp.

Construction labor: Who pays for better databases?
The broad goal is to build databases of diverse race-talk—talk where race is both explicit and implicit. Databases that would promote respectful race-talk in its many forms. In bots, we are aiming for dialogs that are responsive to race, advancing diversity in expression through a wider variety of knowledge bases. Further, we are aiming for databases that better support the recognition of and engagement with discriminatory language and hate-speech. The implication here is that, in building these datasets, they will increase the variety of ways humans and bots talk about race, and capture more of the subtitles in that talk.

Building these types of databases does not require cutting edge research that is currently beyond implementation; it simply requires our resources. In the world of facial recognition, Joy Buolamwini is already tackling this problem by collecting images for a more color diverse facial recognition database [51]. The work of building databases is no small task. We must ensure we do this it ethically. We need to account for the labor and profit involved in constructing databases, in what is often considered menial non-technical labor [45,46,70]. Building these corpora is not simply "an API call away"—a phrase Silberman, Irani, & Ross use to characterize many peoples' notion of workers on mechanical Turk [70]. Likewise, we cannot rely on wholesale automation for database generation either. This strategy will always embed the biases inherent in default, "normal" talk. Without explicitly building databases with diverse representations of language, automatic database generation will, inevitably, be unable to handle the wide contexts conversational agents, like bots, will be accountable to. Better databases require attention to the personal labor contributions necessary to construct them. Workers are a critical part of this system, there is no plurality of databases without them.

Ethically developing a diversity of databases opens up possibilities for handling race and racism in language outside the binary pattern-matching of the blacklist. If we had databases capturing the many types of talk we wish to see more of, we would have a larger volume of text to contrast and combat the surplus of racist, hate-speech in networked conversations. However, building a plurality of databases requires us to interrogate our database practices as much as the collections themselves. We cannot continue to do what is fastest, easiest, and most common. These

practices may produce fast turn-around times for business and research projects; however, they come at an unethical cost that is in direct contrast to the goals of developing more database variety [7,76]. We create new opportunities by investing into an array of racially-conscious databases.

Language Processing: What do chatbots understand as language?

Beyond data repositories, we need to engage with the algorithms that dictate a chatbot's inner workings. How do these algorithms relate to larger structures of race, equity, and power? Algorithms literally define the way a chatbot understands the construct of language. Given that conversing with people is a core goal for chatbots, understanding language—the *medium* of conversation—is essential in this context. Learning to converse with others and learning a medium for conversing are two highly interrelated topics, both algorithmically and theoretically, and they are both entangled in structures of race, equity, and power. These topics, learning and language processing, are often separated into the domains of ML and NLP, respectively. While both ML and NLP are within the domain of AI research, they represent separate, vital lenses through which we build, study, and make sense of chatbots. To inquire into a chatbot's understanding of language, into how a chatbot responds in conversation, we need to immerse ourselves within the worlds of NLP.

Making sense of conversation: How much context does a chatbot have?

So, do humans and chatbots have different understandings of language? If you've chatted with an AI bot lately, like the ones from Microsoft's fleet including Zo (English, USA), Xiaoice (Chinese, China), and Ruuh (English, India) [34], there's a good chance you've been left with a peculiar and distinctive sense. This is one that leaves you both impressed with how well the agent holds up and frustrated with its shortcomings. Consider the conversation we had with Zo in Figure 1. When talking about music, we had mentioned a deep love for the hip-hop group *A Tribe Called Red* from Canada, known for blending hip-hop and First Nations sounds. Zo said they had not heard of the band before so we asked her to look them up. Pictured in this figure is the downfall of our conversation. Things start off well, Zo is able to (enthusiastically) recall the topic of conversation from a few turns prior, a huge feat for a chatbot—coreference resolution, recall of facts from a long, multi-turn conversation, is an unsolved and difficult problem in NLP [55]. However, things devolve quickly after responding to our typo-ed follow-up, asking "what did you (sic) thing?" What might make a bot believe that Choctaw, a tribe from Southeastern North America, is a reasonable response to such a question? If an agent sees language as a set of symbols that have categorical associations, such a bot might determine that the conversation is referencing people from an indigenous nation and then respond with any indigenous tribe name it can recall. However, even if this is a reasonable interpretation from the chatbot's point of view, this paring down of language is *not* a reasonable interpretation from our side of this conversation. In fact, this response is an incredibly problematic, discriminatory utterance, rife with disrespect. So how can we help Zo and other bots do better?

We know that the chatbot doesn't have the context for language that we have, the context that tells you it's racist to respond in a way that flattens the variety of and differences between thousands of indigenous nations into a single name stored to memory. However, just because Zo exists in a silicon space without our context does not mean that the context we bring to a conversation suddenly disappears. The contexts of our worlds are still present, whether a chatbot understands that or not.

Focus on syntax: What role does theory play?

This removal of context is a critical part of NLP's history. Influenced by Chomsky's 1957 publication *Syntactic Structures*, NLP made major advances building off the concept of generative grammars, formalized through context-free grammars [18,67]. Generative grammar as a construct focuses primarily on the *syntactic* aspects of language, mostly bracketing away other sub-domains of linguistics like semantics and pragmatics. While semantics has garnered attention within the world of NLP, pragmatics is an incredibly difficult and under-researched domain.

This matters precisely because of pragmatics, the context and use of language. Conversation is full of pragmatics. Talk is woven with references to things in the world, things we've said before, cultural conventions, and more. While there have been numerous technical and theoretical turns and foci in NLP, including a turn to semantic grammars in the 80s, much is indebted to this exclusive and exclusionary focus on syntax [49,67]. In more recent NLP trends towards probabilistic variants of formalized grammars [e.g., 44], language is constructed through a focus on the ordering of words, the structure of the data—an orientation towards language that is close to syntactic structuring. Ultimately, there are many syntactically valid utterances an AI bot can generate, including a glut of racist context. However, since the predominant focus has been on generating syntactically-valid utterances and valid utterances alone, we are not able to technically contend with the consequences of context at a structural level. We have been deferring the inordinately difficult but ever-present challenges that pragmatics and semantics present. The trouble of the world is always, already in our language, even when we attempt to bracket away complexity. Zo need not understand the "trash heap" of pragmatics to draw from and contribute to the way the world is embroiled in language [26]. Nevertheless, Zo is already acting with its own machine intelligence, from its own position of agency and context [73].

Context and Agency: Can distributed networks of chatbots expand machine intelligence?

If we take a machine's context and agency as a starting point, how can they contend with race in language?

Through an NLP lens, race-talk is difficult for chatbots, in part, because they come to language from a different context than their human counterparts and with different underlying mechanisms. Understanding the structures that impose difficulty in this problem space allows us to direct our focus on how we can craft new and improved ways for chatbots to handle race-talk. Subsequently, a modified theoretical orientation towards human-machine conversation requires us to consider how different types of actors—with very different capacities—come together to co-constitute talk that is collectively meaningful.

While there is a great deal of potential in focusing on this notion of different types of actors with very different capacities, it appears that an underlying assumption of a generalized chatbot like Zo is that bots can have conversations embedded in seemingly "universal" cultural contexts. Even if we were to hold a generalized chatbot up to a human standard, what human can have a conversation on literally any topic, in every context, with anyone? These underlying assumptions are at odds with the issue of semantics and pragmatics. Meaning and context are not universal. These constructs come to make sense through their specific and varied networked relationships. Moreover, there is no reason that the number of agents in a conversation should be limited to two generalized, all-purpose actors. The more we think on it, the less clear the idea of "generalized chit chat" becomes—excepting idle pleasantries, and asking for the time or the weather.

Rather, than striving for the abstract and un-situated notion of a general chatbot or a generalized database of conversational talk, we can think about bots with specialized areas of expertise. Although this heterogeneous version of chatbot design might appear to be a simple idea—certainly there exist many domain specific chatbots—consider how this idea expands as bots develop networked relationships through an ensemble. From this alternate view, there is a world of possibility for what corresponding interactions might look like. An ensemble of chatbots—whose knowledge bases and language styles would effectively embody *differing abilities*—allows us to examine the possibilities for how conversations unfold between distributed yet interconnected actors. Moreover, it gives us a different way to handle the difficulties of language *in situ*, difficulties like race-talk.

Consider a conversation where the bot you are chatting with —or the conversation controller bot—realizes the talk may be slipping outside of their domain. Perhaps, in learning this, the bot defers to a network of other bots to bring in help for continuing the conversation? Here, context emerges from the networked structure of conversation, from how and when other actors are solicited, and from how they participate in multi-party conversation. Here, partial and incomplete forms of talk are a desired outcome from chatbots. Unlike "universal" agents, shortcomings in these ensembles would be opportunities for new agents to participate—shortcomings might be accompanied by meta-data and reason-logs, citing issues like confusing language use or an out-of-domain cultural references. By introducing chatbots with partial, fallible language capacities, we are presented with the potential of a very different realm of "natural language" and interaction design.

Machine Learning: How does a chatbot's agency impact its conversational learning?

While it may seem like learning algorithms can pick up where NLP's foundations fall short, we cannot ignore the active role machine intelligence plays in producing *contexts* and *meanings* in race-talk. This role is enacted through the process of taking an input, processing it according to the architecture deployed (say Long Short-Term Memory, LSTM), and, eventually, converging to produce an output. More specifically, given a dataset, an algorithm infers statistically notable patterns. Critically, these patterns arise through the constraints of the machine's material circumstances, e.g., the nature of the data, the chosen algorithm, the setup of this algorithm, etc. It's in this combination of code and materials that machine learning algorithms build up an internally consistent world; by applying particular modes of operation and logic in a given setting, they come to have an agency of their own in what we can call *world-building*.

Frequently, the predictions enacted in one world (often a small proxy or 'toy' world) are subsequently operationalized in another, machine-external context of operationalization. Two worlds, one contained and rendered in an internally consistent context and the other open and subject to the chaos of real-world applications are treated, for practical purposes, as similar if not the same. So how does this collision of machine-internal world-making and machine-external world-application contribute to the difficulties in handling race-talk? In searching for improved ways to handle race-talk, we need to interrogate this agency of machine learning algorithms, particularly opaque and popular models like neural nets.

From their own internal point of reference, algorithms learn to converse based on the predictions of their world-building, their internal context. Thus, a chatbot that learns, after some *n* iterations, that *Choctaw* is an adequate response to *A Tribe Called Red* has an interior world that rewards the learning of racist associations and of flippant contemplation, like turtles as a non sequitur. If chatbots are to be more responsive to and responsible for inferences like these, it's clear we need better ways of reconciling the differences between machine-internal and machine-external worlds. However, not all algorithms allow us to understand their interior worlds. Neural nets, particularly deep neural nets, have hailed in a wave of high-accuracy prediction in machine learning at the cost of us being able to understand or adjust their internal states. While prediction accuracy is enticing, an algorithm's internal conditions are critical to account for what is learned and how this learning is

actioned, how it comes to have agency. If we're unable to understand the internal world of a machine in its own right, how will we build a deeper understanding of the differences between chatbot and human worlds, and how will we make the differences generative?

Making sense of internal and external worlds: How do our social worlds develop relationships with ML algorithms?

But, because machine learning is embedded in the language of abstraction, it can be incredibly difficult to make sense of how technical and theoretical algorithmic complications connect back to our experiences of the world and to problem spaces like race-talk. While the following example is outside the direct problem space of race-talk, it concretely illustrates how the inner-contexts of algorithms are agentially contributing to machine-external worlds. Starting in the 1990s, machine learning algorithms have been studied and deployed for predicting the risk of pneumonia in a healthcare context [15,22,23]. As explained by Caruana et al. in a 2015 publication, the goal of these studies is to predict the probability of death in order to improve the chances that high-risk patients would receive better care [15]. These studies compare the outcomes of a number of machine learning models, including a rule-based model and a neural net model. Unsurprisingly, the neural net was the most accurate model; it did the best. But it was ultimately deemed too dangerous to use with actual patients. Accuracy is not necessarily the best measure for evaluating a machine learning algorithm. Now, this can seem counter intuitive—especially because we rely so heavily on accuracy to understand if a model is doing well. But accuracy cannot tell you when your algorithm has learned that patients with asthma are low-risk, despite the fact that healthcare professionals know pneumonia patients with asthma are in incredibly high-risk. Within the internal-context of the algorithm, asthma patients did not die of pneumonia frequently and so they were deemed to be low risk. The algorithm had no way to account for the external fact that these patients were always hospitalized because of their high-risk status, which is why so few patients with asthma died of pneumonia. Despite the abstraction, there are specificities of the machine-external world—the context—which pose problems for machine learning algorithms, especially low-interpretability high-accuracy algorithms like neural nets.

Accounting for race: What are some of the ways that race becomes situated within algorithmic agency?

These problems relate to race as well, both inside and outside of healthcare. In the world of United States healthcare, there is empirical evidence that black people receive inadequate treatment recommendations for pain management [41]. A substantial number of white medical students and residents held unfounded racist beliefs about how much pain black people experience, which led to *recommending less treatment* for black patients than for white patients. It is highly likely that patients are receiving racially biased treatment recommendations. As a result, there may be bias in the patient records around the country, reflected in data features like the dosages black patients received for pain medication. What happens if a hospital wants to use patient records in an algorithm that helps practitioners determine treatment outcomes, like medication dosage levels? What do we do in hospital settings that have already incorporated these systems into their practitioner work practice [24]? How do we account for this type of bias when developing and deploying virtual nursing bots [11]? Outside of healthcare, machine-external entanglements with race have major implications for an algorithms agency. Amazon developed an algorithm that perpetuated discriminatory redlining practices, rolling out one-day Prime shipping almost exclusively to white neighborhoods in major US cities by focusing on zip-codes with high-density prime memberships [44]. Amazon's algorithm did not contend with race directly in the machine's internal context, Amazon stated that race was not even a part of the algorithm. But blacklisting race did not stop the propagation of discriminatory practices. These entanglements come up in language as well. Google's advertising algorithms, AdWords and AdSense, delivered discriminatory advertisements in search results for black-identifying names [72]. Based on the name alone, Google was more likely to generate ads suggesting the person being searched had been arrested for black-identifying names. Algorithms are agential. They are working within networked social and technical systems in ways that engages with the structures of race and race-talk.

Reconsidering how we build and evaluate ML: Are we asking enough of ourselves? Enough of the algorithms?

Just because an algorithm has a high accuracy, does not mean what it is learning is right, optimal, or ethical. It is simply a reflection of a machine making use of its learning algorithm to discover patterns in the data. And while some people may say you just need a better dataset, we still need to learn to work with the data available. There is no perfect dataset. As noted by the authors of the pneumonia study, "[learning] must be done with the data that is available, not the data one would want" [15].

Working with the world as it is now, with the data that exists, is key to algorithmic accountability. Professional dialogue on becoming more responsible for the agency of algorithms frequently focuses on setting up key guiding principles, taking a nod from previous U.S. policy setting [1,27,63,74]. A lot of this dialogue focuses on fairness and transparency, but there is good reason to ask if these visions for algorithmic accountability go far enough. In particular, there is a conflation that being able to see what is happening within a system, i.e., transparency, and making a system accountable [6]. Knowing that an algorithm is contributing to racial bias does not go far enough in addressing the social and technical components that enable this reality. It does not make us accountable. So, how do we move forward in a way that enables us to concretely develop accountable, response-able algorithms?

Interpretability and Tunability: What agency comes out of making sense of an algorithm?

These questions of algorithmic accountability are especially difficult in the context of neural nets, which are just beginning to garner research output in this domain. In general, interpretability, sometimes used synonymously with transparency, plays a key role in the technosocial networks of algorithmic accountability. In the case of the pneumonia risk research, it was the more interpretable algorithms, like the rule-based model, that revealed the learning of the deadly, externally-incorrect asthma association [15]. The more interpretable models of the pneumonia study allowed for adjustments to counter-act dangerous, problematic learning. On the other hand, neural nets—while often lauded as magically accurate—pose serious problems for adjustment. Crucially, the possibility of adjustment allows for response-ability. Even though there are some techniques for "repairing" neural nets, these techniques frequently require removing problematic data, further constraining the machine-internal context—and paralleling the repair work of the blacklist [15]. The reality is that neural networks aren't going anywhere. While advances in machine learning have resulted in high-accuracy, interpretable, and adjustable models that are a good-fit for healthcare datasets, these models do not address the inscrutability problem of neural nets, nor do they fit text-based datasets well, the kinds that might be used for AI chatbots [15]. Neural nets have also been a key player in recent advances in NLP, advances that play an essential role in the progress of AI chatbots. However, due to their inscrutability, this also poses challenges for conceiving of new and improved ways to handle race talk with neural nets.

There is growing research on interpretability of neural networks, but interpretability of a neural net does not necessarily mean there is room for adjustability [5,57]. While encouraging, much of this research shares underlying assumptions with the transparency value that Ananny and Crawford deeply trouble [6]. In light of these elements, what happens when there are problems with a neural net we don't know about and have no way of adjusting?

We need neural nets that are tunable. Nets (and ensembles) that can be adjusted and response-able to their technosocial networked context, a context entangled with machine-internal and machine-external consequences. Transparency is severely limited (Ananny 2016). We whole-heartedly agree that asking for transparency—or its stand-in, interpretability—is not enough. However, striving for tunable neural nets may fundamentally disrupt their black box abstraction. When thinking about how these types of models can be tunable, we need to examine the ways neural networks are already being adjusted and modified. Developing a technical, practicable notion of tunability requires in depth investigations—basic research—into the ways these networks are already being tuned through things like pre-training [33], initial weight setting (like Xavier initialization), controlling ensembles of recurrent neural networks in real-time [4], and systems that have emphasized refinability of deep neural nets [47]. We are already participating in finicky behaviors with neural nets to help them converge or produce "optimal" outputs. When we have more established understanding of how we can work with these nets to tune and refine their outputs, we can push ourselves further in exploring how to tune neural nets and other deep learning models to be more response-able to network of worlds they participate in.

Interdisciplinary Partnerships: How we can leverage, cross-domain collaborations built on advocacy?

Returning to our opening question, how does a chatbot's agency impact its conversational learning, we are confronted with one more essential question, how do we best understand a chatbot's agency? Best perform the basic research into response-able deep learning? When we consider the large technosocial networks that these deep learning models are embedded within, we come face to face with the various domain specific worlds that need to be understood to more fully make sense of, reflect on, and evaluate these deep learning algorithms. There is an urgency to study algorithms that are already in use [12,19] and to study the entire development cycle for generating deep learning algorithms. If we take seriously the challenge of tunable deep learning models, we must also critically interrogate how we pick problems in non-ML domains, understand when an output "looks right," and evaluate what exactly the contribution of the output is in other fields—where it fits within a fields historical and contemporary knowledge. When taking seriously the knowledge domain of worlds outside of machine learning alone, we can come to novel and challenging interpretations of a system's output and its implications. Leahu gives us a glimpse of the power of non-normative interpretation by providing a relational perspective on the agency of learning algorithms [54]. If we set out to endeavor into deep intellectual inter-cultural exchanges—home-stays if you will—with intellectuals from other domains, we are opening up the possibility of building algorithms in a more deeply connected, networked technosocial ecosystems. In this type of ecosystem, we will be better able to address the concerns of algorithmic accountability, and build futures that value and embody the plurality of worlds that exist. These types of long-form collaborations are vital for developing and understanding the agency of a chatbot and its relationship to larger social systems like race and race-talk.

WHERE DO WE GO FROM HERE?

While this work covers quite a bit of ground, this is only one step in a much larger problem space. In this paper, we have outlined a program that we as a community need to undertake in order to create chatbots capable of more than simply cutting out words. Critically, in taking on Donna Haraway's call to "stay with the trouble" we are holding onto an understanding of the worlds we have cut through that will always confront us with complexities—where

critical and profoundly important issues cannot be addressed through neat separations between what people do and how machines, like chatbots, operate. In determining where we go from here, it is important that we hold onto the complexities of our lived experiences and refuse to reduce human struggle into something that is uniform or singular.

Building Better Worlds

As we chip away at the foundational tensions uncovered throughout this paper, we are striving to enable futures in which chatbots are better able to handle the complexities of race-talk. Born out of the tensions of this problem space is an understanding that any building of better worlds requires a foundation in the *troubles* and in continually seeking a route that prioritizes equity, justice, diversity, difference, and respect. Crucially, this call for building better worlds is not a utopian vision founded on the myth of a universally perfect future. Rather, this paper explores a way of living and striving for change founded on technosocial networks that are always, already fragile and chaotic. Even though the technologies involved in developing AI chatbots start from a place of struggle in relation to race, the promise of something better rests on the recognition that we are always in the process of making and unraveling our worlds.

Not surprisingly, chatbots, and conversational agents more broadly, are already being employed to build better worlds, both in research and in industry. There have been a range of studies involving conversational agents presented at CHI that have artificial agents involved in nursing, educational settings, activism, and conflict resolution [58,66,68,80,81]. Much of this work is focused on the interaction interface, studying things like the impacts of various avatars on embodied agents. While not all work is invested in imitating or replicating humanistic qualities, a notable portion of this work is focused on if agents can achieve human-like abilities through talk and embodied presence.

Putting to one side this question of replicating human capacities—so thoroughly debated and contested in ongoing conversations surrounding the Turing Test [36,75]—our concern has centered on the ways that technological artifacts like bots have politics [79]. Whether or not race has been actively accounted for, artificial agents are already implicated in the structures of identity and race. Even if bots do not currently have the technical capacities to handle race well, we have struggled through the technological structures at play in order to explore alternative and just possibilities for tech that is more responsive to race-talk and racism. *Taking-the-technology-seriously* has been central to our work. Through it, we have worked to ground problems and tensions and to show how politics of/with race intersects and entangles with the technical. Essentially, we've exposed how race comes to matter, and where the material conditions of possibility might lay for making a difference and building better worlds.

As it stands, among these always emerging, technosocial networked relations where race is ever-present, the question left to ask is *what racial affinity is your chatbot?*

REFLEXIVE DISCLOSURE: RECOGNIZING OUR ROLE

Race is a distributed, global system that we are all implicated in. When it comes to the design of chatbots—and human-machine interactions more generally—we must acknowledge our complicity in the worlds we are making.

No matter where you live, race makes an impact on your life. The unfortunate reality is that for those with privileged racial identities, it can be easy—normal—to lose sight of how race is impacting your experiences in the world. If you find yourself coming to the realization that you had not thought much about race in the past, it is likely that you are benefiting from racial privilege. As such, it is critical that everyone step up and engage in practices that address the complexities of race head on. There are important voices that are absent from this work. The identities of the authors only represent a small and privileged subset of racial identities. To ensure our voices are just part of a much larger dialog happening in this space, we make space for voices that are different than our own throughout this piece. Further, it would be an outright lie to say that we, the authors, are outside of racism. When we acknowledge our racism, it allows us to identify problematic systems and behaviors and then inhibit them. We take a stand against racism because addressing this problem directly is the only way that we all can work on reducing the impact of racism.

CONCLUSION: HOW DO WE EMBRACE THE TROUBLE?

Talking about race is not easy, for people or bots. Conversations about race expose our identities, our affinities, and our politics. Through their relationship to history and culture, conversations lay bare how bias is built into our actions, our language, and the technologies we live with and through. Whether we want to engage or not, race is entangled with our world and our conversations—conversations that chatbots are a part of

In writing this paper, we set two essential questions to guide this work 1) How can chatbots handle race in dialog in new and improved ways? and 2) Why is race-talk so difficult for chatbots? Given the complexities in the problem space of race and chatbots, these questions unraveled into many narrower, domain-specific question as we worked through the technologies in each domain. Stepping through the challenges of investigating databases, natural language processing, and machine learning in conjunction with critical, intersectional theories, essential questions have helped guide our inquiry. These questions open up possibilities for creating new technosocial contexts connecting people and machines. Through making tangible the abstract and disparate qualities involved in working with race and chatbots, this paper works as a synthetic guide, pointing us towards the pursuit of possible futures where chatbots are intentionally more capable of handling race-talk in its many forms.

ACKNOWLEDGMENTS
Removed for blind review.

REFERENCES

1. ACM US Policy Council. 2017. *Statement on Algorithmic Transparency and Accountability*.
2. Sara Ahmed. 2004. Declarations of Whiteness: The Non-Performativity of Anti-Racism. *borderlands e-journal* 3, 2.
3. Sara Ahmed. 2017. *Living a Feminist Life*. Duke University Press#.
4. Memo Akten and Mick Grierson. 2016. Real-time interactive sequence generation and control with Recurrent Neural Network ensembles. *Neural Information Processing Systems 2016*.
5. David Alvarez-melis and Tommi S Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *EMNLP 2017*.
6. Mike Ananny and Kate Crawford. 2016. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*: 1–17. http://doi.org/10.1177/1461444816676645
7. Alyx Baldwin. 2016. The Hidden Dangers of AI for Queer and Trans People. *Model View Culture*. Retrieved May 25, 2017 from https://modelviewculture.com/pieces/the-hidden-dangers-of-ai-for-queer-and-trans-people
8. James Baldwin. 2010. As Much Truth as One Can Bear. In *The Cross of Redemption: Uncollected Writings*, Randall Kenan (ed.). Pantheon Books, New York.
9. Karen Barad. 2007. *Meeting the Universe Halfway*. Duke University Press, Durham.
10. Jamie Bartlett. 2015. A Life Ruin: Inside the Digital Underworld. *Medium*. Retrieved September 16, 2017 from https://medium.com/@PRHDigital/a-life-ruin-inside-the-digital-underworld-590a23b14981
11. Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. 2009. Taking the Time to Care: Empowering Low Health Literacy Hospital Patients with Virtual Nurse Agents. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1265–1274. http://doi.org/10.1145/1518701.1518891
12. Danah Boyd and Kate Crawford. 2012. Critical Questions for Big Data. *Information, Communication & Society* 15, 5: 662–679. http://doi.org/10.1080/1369118X.2012.678878
13. Peter Bright. 2016. Tay, the neo-Nazi millennial chatbot, gets autopsied. *Ars Technica*. Retrieved August 27, 2017 from https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/
14. Aylin Caliskan-Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *Science* 186, April: 183–186. http://doi.org/10.1126/science.aal4230
15. Rich Caruana, Paul Koch, Yin Lou, Johannes Gehrke, and Marc Sturm. 2015. Intelligible Models for HealthCare : Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM. http://doi.org/http://dx.doi.org/10.1145/2783258.2788613
16. Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #Thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ACM, 1201–1213. http://doi.org/10.1145/2818048.2819963
17. Ethan Chiel. 2016. Who turned Microsoft's chatbot racist? Surprise, it was 4chan and 8chan. *Splinter News*. Retrieved September 16, 2017 from http://splinternews.com/who-turned-microsofts-chatbot-racist-surprise-it-was-1793855848
18. Noam Chomsky. 2002. *Syntactic Structures*. Mouton de Gruyter, Berlin.
19. Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*: 1–14. http://doi.org/10.1177/2053951717718855
20. Rodney Coates. 2007. Covert Racism in the U.S. and Globally. *Sociology Compass* 2, 1: 208231. http://doi.org/10.1111/j.17519020.2007.00057.x
21. Beth Coleman. 2009. Race as Technology. *Camera Obscura* 24, 1.
22. Gregory F Cooper, Vijoy Abraham, Constantin F Aliferis, et al. 2005. Predicting dire outcomes of patients with community acquired pneumonia. 38: 347–366. http://doi.org/10.1016/j.jbi.2005.02.005
23. Gregory F Cooper, Constantin F Aliferis, Richard Ambrosino, and John Aronis. 1997. An Evaluation of Machine-Learning Methods for Predicting Pneumonia Mortality.
24. Kate Crawford and Ryan Calo. 2016. There is a Blind Spot in AI Research. *Nature* 538, 7625: 311–313.

http://doi.org/10.1038/538311a

25. Kimberle Crenshaw. 1991. Mapping the Margins: Intersetionality, Identity Politics, and Violence Against Women of Color. *Stanford Law Review* 43, 6: 1241–1299.
26. Gilles Deleuze and Felix Guattari. 1987. *A Thousand Plateaus: Capitalism and Schizophrenia*. University of Minessota Press, Minneapolis.
27. Nicholas Diakopoulos. 2014. *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*. Columbia University Academic Commons. http://doi.org/10.7916/D8ZK5TW2
28. Tawanna R Dillahunt. 2014. Fostering Social Capital in Economically Distressed Communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 531–540. http://doi.org/10.1145/2556288.2557123
29. Paul Dourish. 2014. No SQL: The Shifting Materialities of Database Technology. *Computational Culture*, 4: 1–37.
30. Jacob Eisenstein. 2013. What to do about bad language on the internet. *Naacl-Hlt*, Association for Computational Linguistics, 359–369.
31. Sheena Erete and Jennifer O Burrell. 2017. Empowered Participation: How Citizens Use Technology in Local Governance. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, 2307–2319. http://doi.org/10.1145/3025453.3025996
32. Sheena L Erete. 2015. Engaging Around Neighborhood Issues. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*: 1590–1601. http://doi.org/10.1145/2675133.2675182
33. Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research* 11, Feb: 625–660.
34. Mary Jo Foley. 2017. Microsoft launches Ruuh, yet another AI chatbot. *ZDNet*. Retrieved September 4, 2017 from http://www.zdnet.com/article/microsoft-launches-ruuh-yet-another-ai-chatbot/
35. Andrea Grimes, Martin Bednar, Jay David Bolter, and Rebecca E Grinter. 2008. EatWell: Sharing Nutrition-related Memories in a Low-income Community. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, ACM, 87–96. http://doi.org/10.1145/1460563.1460579
36. Barbara J Grosz. 2012. What Question Would Turing Pose Today? *AI Magazine* 33, 4: 73–81. http://doi.org/10.1609/aimag.v33i4.2441
37. David Hankerson, Andrea R Marshall, Jennifer Booker, Houda El Mimouni, Imani Walker, and Jennifer A Rode. 2016. Does Technology Have Race? *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ACM, 473–486. http://doi.org/10.1145/2851581.2892578
38. Donna Haraway. 1991. A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. In *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, New York, 149–181.
39. Donna J. Haraway. 1991. *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, New York. http://doi.org/10.2307/2076334
40. Donna J. Haraway. 2016. *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press, Durham.
41. Kelly M. Hoffman, Sophie Trawalter, Jordan R. Axt, and M. Norman Oliver. 2016. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences* 113, 16: 4296–4301. http://doi.org/10.1073/pnas.1516047113
42. bell hooks. 2003. Talking Race and Racism. In *Teaching Community: A Pedagogy of HOpe*. Routledge, New York, NY, 25–40.
43. Helena Horton. 2016. Microsoft deletes "teen girl" AI after it became a Hitler-loving sex robot within 24 hours. *The Telegraph*. Retrieved August 27, 2017 from http://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/
44. David Ingold and Spencer Soper. 2016. Amazon Doesn't Consider the Race of Its Customers. Should it? *Bloomberg*.
45. Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 611–620. http://doi.org/10.1145/2470654.2470742
46. Lilly C Irani and M Six Silberman. 2016. Stories We Tell About Labor: Turkopticon and the Trouble with "Design." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, 4573–4586. http://doi.org/10.1145/2858036.2858592
47. Natasha Jaques, Shixiang Gu, Richard E Turner, and Douglas Eck. 2017. Tuning Recurrent Neural Networks with Reinforcement Learning. *ICLR Workshop*.

48. Sarah Jeong. 2016. How to Make a Bot That Isn't Racist. *Motherboard*. Retrieved May 25, 2017 from https://motherboard.vice.com/en_us/article/how-to-make-a-not-racist-bot

49. Daniel S Jurafsky and James H Martin. 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. http://doi.org/10.1162/089120100750105975

50. Darius (Dariusk) Kazemi. 2016. wordfilter. *npm*. Retrieved May 30, 2017 from https://www.npmjs.com/package/wordfilter

51. Zoe Kleinman. 2017. Artificial intelligence: How to avoid racist algorithms. *BBC News*.

52. David Kushner. 2015. 4chan's Overlord Christopher Poole Reveals Why He Walked Away. *Rolling Stone*. Retrieved September 16, 2017 from http://www.rollingstone.com/culture/features/4chans-overlord-christopher-poole-reveals-why-he-walked-away-20150313

53. Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*.

54. Lucian Leahu. 2016. Ontological Surprises: A Relational Perspective on Machine Learning. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, ACM, 182–186. http://doi.org/10.1145/2901790.2901840

55. Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's Multi-pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, Association for Computational Linguistics, 28–34.

56. Peter Lee. 2016. Learning from Tay's introduction. *Official Microsoft Blog*. Retrieved June 1, 2017 from https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.0000fpjmog51cfpxpwz11olji2ndk

57. Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 107–117.

58. Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*: 5286–5297. http://doi.org/10.1145/2858036.2858288

59. Tara McPherson. 2011. US Operating Systems at Mid-Century: The Intertwining of Race and UNIX. *Race After the Internet*. http://doi.org/10.4324/9780203875063

60. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ICLR Workshop*.

61. Lisa Nakamura. 1995. Race In/For Cyberspace: Identity Tourism and Racial Passing on the Internet. *Works and Days* 13: 181–193.

62. Gloria Naylor. 1986. The Meanings of a Word.

63. Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, et al. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. *FAT/ML*. Retrieved June 15, 2017 from http://www.fatml.org/resources/principles-for-accountable-algorithms

64. Sarah Perez. 2016. Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism. *Tech Crunch*. Retrieved August 27, 2017 from https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/

65. Derek Powazek. 2013. What online communities can learn from twitter's "block" blunder. *Wired Magazine*. Retrieved June 5, 2017 from https://www.wired.com/2013/12/twitter-blocking-policy/

66. Emilee Rader, Margaret Echelbarger, and Justine Cassell. 2011. Brick by Brick: Iterating Interventions to Bridge the Achievement Gap with Virtual Peers. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2971–2974. http://doi.org/10.1145/1978942.1979382

67. Stuart Russell and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, New Jersey. http://doi.org/10.1016/0925-2312(95)90020-9

68. Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. 2016. Botivist: Calling Volunteers to Action Using Online Bots. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ACM, 813–822. http://doi.org/10.1145/2818048.2819985

69. Ari Schlesinger, W Keith Edwards, and Rebecca E Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, ACM Press, 5412–

5427. http://doi.org/10.1145/3025453.3025766

70. M Six Silberman, Lilly Irani, and Joel Ross. 2010. Ethics and Tactics of Professional Crowdwork. *XRDS* 17, 2: 39–43. http://doi.org/10.1145/1869086.1869100

s-tay-is-an-example-of-bad-design-d4e65bb2569f

72. Latanya Sweeney. Discrimination in Online Ad Delivery.

73. Alex S Taylor. 2009. Machine Intelligence. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2109–2118. http://doi.org/10.1145/1518701.1519022

74. Zeynep Tufekci. 2015. Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency. *Journal of Telecommunications and High Technology Law* 90: 203–218.

75. Alan M Turing. 1950. Computing machinery and intelligence. *Mind* 59, 236: 433–460.

76. James Vincent. 2017. Transgender YouTubers had their videos grabbed to train facial recognition software. *The Verge*.

77. Kevine A. Whitehead. 2009. "Categorizing the Categorizer": The Management of Racial Common Sense in Interaction. *Social Psychology Quarterly* 72, 4: 325–342.

78. Keving A. Whitehead and Gene H. Lerner. 2009. When are persons "white"?: on some practical

71. Caroline Sinders. 2016. Microsoft's Tay is an Example of Bad Design. *Medium*. Retrieved August 27, 2017 from https://medium.com/@carolinesinders/microsoft-

asymmetries of racial reference in talk-in- interaction. *Discourse & Society* 20, 5: 613–641. http://doi.org/10.1177/0306312706069437

79. Langdon Winner. 1980. Do Artifacts Have Politics? *Daedalus* 109, 1: 121–136.

80. Jun Xiao, John Stasko, and Richard Catrambone. 2007. The Role of Choice and Customization on Users' Interaction with Embodied Conversational Agents: Effects on Perception and Performance. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1293–1302. http://doi.org/10.1145/1240624.1240820

81. Qianli Xu, Liyuan Li, and Gang Wang. 2013. Designing Engagement-aware Agents for Multiparty Conversations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2233–2242. http://doi.org/10.1145/2470654.2481308

82. 2016. Tay AI. *Know Your Meme*. Retrieved June 1, 2017 from https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.0000fpjmog51cfpxpwz11olji2ndk