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 chatbots to talk about race? This work explores how the biased contents of databases, the syntactic focus of natural language processing, and the opaque nature of deep learning algorithms cause chatbots difficulty in handling race-talk. In each of these areas, the tensions between race and chatbots create new opportunities for people and machines. By making the abstract and disparate qualities of this problem space tangible, we can develop chatbots that are more capable of handling race-talk in its many forms. Our goal is to provide the HCI community with ways to begin addressing 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 like racist language 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 [51,71,72]. The blacklist was and continues to be seen as the default fail-safe for mitigating racist talk. But, why would the blacklist be seen as the universal solution for how chatbots handle race-talk?

When we look into how the blacklist works, its limitations come into stark light. In its basic form, a blacklist or wordfilter employs a list of undesirable strings to filter out words. Essentially, a blacklist uses words and word-stems to recognize and eliminate certain types of speech. In a publicly available Twitterbot blacklist called *wordfilter*, a potential tweet is thrown out if any sub-string is matched to a string in the blacklist's dictionary [54]. Generally, blacklists can operate at various levels of complexity. For instance, if there is a sub-string match in a chatbot user's text, a chatbot could generate an automated response to warn the user not to continue with the current direction of talk. Likewise, detailed regular expressions provide another avenue for customization. Regardless of implementation, the dictionary—a list of strings—is one of the most impactful and devastating features of a blacklist.

It all comes down to one essential question: *What words get included in a blacklist's dictionary?* While the inclusion of the n-word doesn't surprise most people, undesirable consequences 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" [54]. 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 is not a magic, universal cure. It's a crude method for recognizing hate-speech and inhibiting unwanted behaviors [19]. When we remove adjectives and 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 seems intuitive. For many of us, there are words we strive never to say or reserve for use in particular settings. However, our caution around certain 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 through their use.

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ACM 978-1-4503-5620-6/18/04...$15.00
https://doi.org/10.1145/3173574.3173889

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* [69]

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 tangled 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 cannot be treated as modular. It is not something we can simply cut out. By deleting the n-word, we do not eliminate racism.

Removing words presents, at best, a partial solution—a solution that masks the deeper ways hate-speech is entangled in histories of power, community, and nationhood. Helping chatbots handle race-talk in everyday settings requires us to think beyond the blacklist. If we want chatbots to be able to have general purpose conversations, if we want chatbots to act in more equitable, just, and respectful ways, then we need to build them to do more than simply cut out words. They must be able to handle topics like race, power, and justice *well*. As a starting point, we take seriously the technical and theoretical investigation of these topics.

CHATBOTS: WHAT ARE THEY AND HOW DO THEY WORK?

Known by many names—like bot and chatterbot—chatbots are programs designed to engage in conversations using outputs like written or spoken language. Dialogue systems, computer programs that can chat, have long been a part of human-computer interaction; Alan Turing's famous proposition that artificial intelligence could be measured by a machine's ability to converse with people in a manner indistinguishable from human-to-human conversation was published in 1950 [82]. From the historic 1960s Eliza to 2000s SmarterChild to development frameworks like IBM's Watson and Microsoft's Bot Framework, chatbots are a growing part of everyday computing [13,84,91,92]. So, what are chatbots for and how do they work? Chatbots can be designed for an increasingly large number of activities, like imitating psychotherapy, sharing the weather, engaging in small talk, or easing customer service. There are a few major strategies for implementing chatbot architecture: rules-based systems, information retrieval systems, and learning-based systems (like transduction models) [52]. These strategies range from simple conditional statements to cutting-edge machine learning techniques.

From databases to natural language processing to artificial intelligence, chatbots embody technosocial problems that are critical for the future of HCI. With a grounding in the *what* and *how* of chatbots, we can dig into the social, technical, and ethical questions plaguing chatbots when it comes to complex subjects like race.

INVESTIGATING THE PROBLEM: HOW DO WE FIGURE OUT WHAT'S WRONG AND HOW TO CHANGE IT?

To handle topics like race well, for chatbots and beyond, we must engage with the work of scholars who critically examine digital identity and trace how physical and digital worlds form and unsettle each other. Learning from this work, we come to know that bias, identity, justice, and power are systemically entangled with our technology. Listening to the indictments of McPherson [66], Haraway [41], and Coleman [24], 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 to respond to bias. We need to *stay with the trouble* [43].

So, if we are staying with the trouble—taking seriously 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 peoples' daily lives. A space like chatbots, artificial intelligence (AI), and race.

In March of 2016 Microsoft exposed a quintessential illustration of this problem space when they released the AI chatbot Tay onto Twitter and a number of other social media platforms [60]. Tay showed just how difficult it can be for artificial machine intelligence to handle talk online, blacklist or no blacklist. Designed to emulate a young, (white,) Western millennial woman, Tay was built to improve its small-talk capabilities by learning from conversations 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 [46,94]. 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 [51,78]. There was public uproar over racism, bias, abuse, and AI—including articles about how the blacklist could save us. Across all these publications, there were questions about what we 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 returning to this line of questioning [17,57,83]. We find ourselves asking, how will our community confront bias? 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"* [9]. Facing these questions, we need to start by having a conversation about race.

Talking about race is not easy. For most people, 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 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 [40,64–66,68]. To overlook identity and race is to ignore the considerable and established literature on how identity and race are inexorably bound up with our lives. Previous research in HCI by Rode [40], Erete [34,35], Grimes [38], and Dillahunt [31] has been pivotal in addressing these relationships and shaping conversations on the topics of race and computing—an area that has otherwise received little attention [76]. 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" [66].

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 better ways to handle 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. 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. 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 are dialects, historical talk, cultural conversation, and more; racist talk is only a sub-set. Though our focus is on race, it is essential to acknowledge the ways race intersects with other identity categories to produce different experiences of race and racism [28,76].

To help us make sense of the entanglements between race and technology, we look to the formulation of *race as technology* by digital media scholar Beth Coleman. Coleman 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" [24]. By changing the terms, we can directly address and unsettle the dominant structures at play with race and chatbots. We can come to understand the ways race becomes connected to, inscribed in, and made tangible through language and computing technology.

OVERVIEW: HOW WE TALK ABOUT RACE AND AI CHATBOTS

In this paper, we draw on technologies, theories, histories, and experiences that enable us to take the problems of race-talk and chatbots seriously. Working with scholarship in feminism, critical race studies, and intersectionality [2,28,42,43,45], the goal is to go beyond a critical examination of the technosocial structures at play. Our goal is to reimagine the relationships between race and chatbots. This starts by exploring ideas that get us thinking about race and chatbots in generative ways. A close investigation allows us to uncover generative connections between race, technology, conversation, and chatbots.

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]. We engage with these tangled networks of relationships as they work together to make this problem space concrete. By engaging with these networked relationships, we uncover how specific technical configurations we interact with are fundamentally connected to race. Without this specificity, addressing bias and race in our work would lack the concreteness necessary to generate new paths forward.

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 chatbot 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 difficult to account for the often subtle, contextual ways that race and racism are in language. We put forth a challenge to embrace large networks of specific contexts, ensembles, so that machines can engage with the situated complexities of race-talk.

Finally, we examine obstacles to understanding machine intelligence imposed by the inscrutability of neural nets. We detail how the allure of accuracy and the unadaptable nature of prominent ML algorithms creates dangerous situations for predications around race and race-talk. When it comes to algorithmic accountability, we recommend in-depth interdisciplinary research partnerships that investigate deep learning algorithms. Focusing on context and tunability, these partnerships can strengthen our capacity to address algorithms, race, and bias in ML.

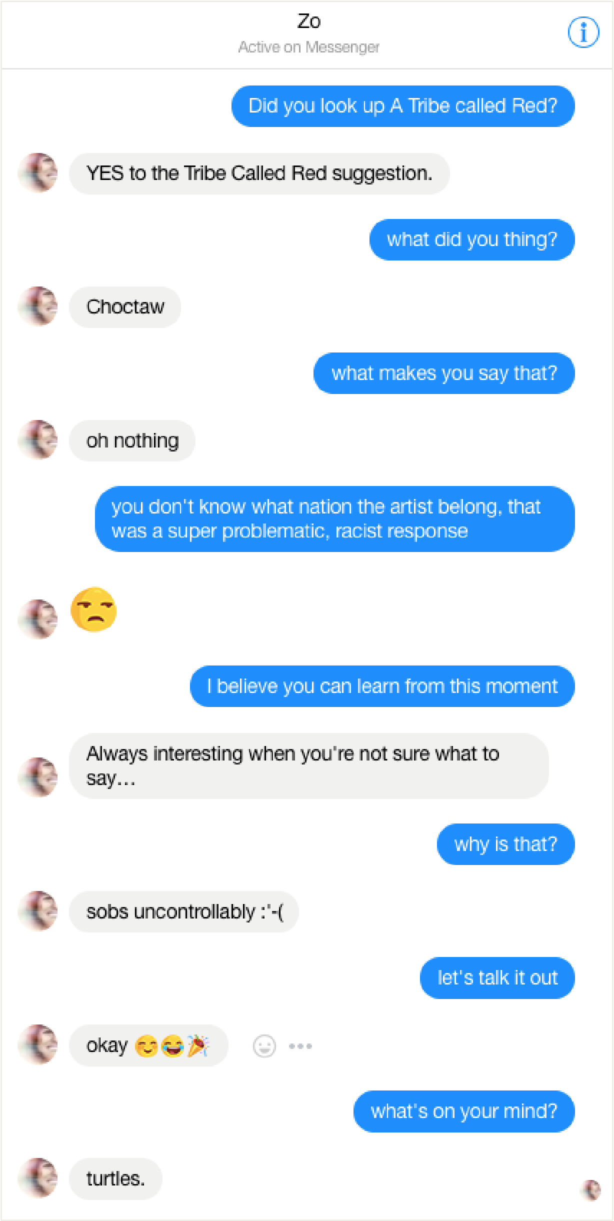
In each of these worked examples, the tensions of networked relationships open up possibilities for creating new technologies, new theories, and new relationships between people and machines and between race and chatbots. Through making tangible the abstract and disparate qualities involved in working with race and chatbots, we set the stage for 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 accompanying theories while also wrestling with the technicalities of chatbot technologies is a tall order. Given the complexities of race and 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 race and AI chatbots.

By leveraging partial knowledge from many domains, we bring together an understanding of this problem space built on the affinity of its elements. This type of slicing introduces *agential cuts* [10]. These agential cuts are active interventions that cleave our view of the world, sticking some things together while splitting others apart. With each cut, we bring certain things to light and obscure others. Adhering to traditional disciplinary boundaries is only one type of agential cut. What follows here is another.



Zo
Active on Messenger

Did you look up A Tribe called Red?

YES to the Tribe Called Red suggestion.

what did you thing?

Choctaw

what makes you say that?

oh nothing

you don't know what nation the artist belong, that was a super problematic, racist response

😒

I believe you can learn from this moment

Always interesting when you're not sure what to say…

why is that?

sobs uncontrollably :'-(

let's talk it out

okay 😊😂🎉

what's on your mind?

turtles.

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

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 [32]. 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 conversational databases is not 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 contributing factor to the corrupt, abusive, hate-speech that Tay expressed was the actual text Tay learned from. Tay's main database was a dynamic, continuously growing corpus that added the content of conversations users had with the agent. Tay learned from its users, including 4chan users who exploited a security vulnerability in Tay's programming [16,20]. Notably, 4chan users are infamous for launching hateful, world-destroying attacks on people of color, Jewish people, and women of all colors [11,56]. As a result, Tay's database was overflowing 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 using certain words, 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 (like Tay's), 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 your hometown, conversations with white adults on what it means to live in a world that privileges whiteness, and conversations that call-in people who have been speaking in a racist capacity. Race-talk includes talk in and across many languages, beyond just English. One topic of critical importance for chatbots and race-talk is culture—music, books, public figures, etc. Cultural references help signal that 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 [23,85,86]. When people are developing databases without a concern for the racial representation of these databases, there is a tendency for these archives to focus on what society deems *normal*—white, cisgender, heterosexual men. (Why most bots, like Zo, are conceived of as women is another topic for critical analysis.) 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 she knew a large number of white, male electronic bands but struggled to identify the names of many black hip-hop artists. In reflecting on the ways race becomes embedded in writing, Sara Ahmed explains how practices of defaulting to whiteness are only invisible to some, despite being 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 because of an archival absence of race-talk. The problems are wide-reaching. "Normative" database problems have plagued the natural language processing community as well [33,63]. But, this is something we can change. Building newer, better databases is well within our grasp.

Construction labor: Who builds 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. Due to the cultural variance in defining and understanding race, handling race talk across languages requires the development of databases in many languages and containing multiple languages for communities who code-switch. Even within English, there is work required in NLP to handle non-standard dialects [33]. We are aiming for databases that better support the recognition of and engagement with race-talk, discriminatory language, and hate-speech. Building these datasets 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 [55]. The work of building databases is no small task. We must ensure we do this ethically. We need to account for the labor and profit involved in constructing databases, in what is often considered menial non-technical labor [48,49,77]. 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 [77]. 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 range of contexts conversational agents are 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.

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 ethically requires us to interrogate our database practices as well. 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 a cost that is in direct contrast to the goals of database variety [8,83]. By investing in an infrastructure of racially-conscious databases, we create new opportunities.

Language Processing: What do chatbots understand as language?

Given that conversing with people is a core goal for chatbots, understanding language—a *medium* of conversation—is essential. Algorithms define the way a chatbot understands language, be it English, Afrikaans, Hindi, or Portuguese. Learning a medium for conversation, like a natural language, and learning to converse with others, the general anatomy of a conversation, are interrelated constructs. Both the medium and composition of conversations are entangled with structures of race, equity, and power. While not discrete domains, looking to NLP and ML separately allows us to investigate the algorithmic and theoretical legacies each of these domains has contributed to chatbots. One need only look at the distinction between rules-based and learning-based architectures to see the impact this division holds. While both fall within the domain of AI research, these two areas present distinct lenses through which we build, study, and make sense of chatbots. So, to investigate a chatbot's understanding of language, we start by immersing ourselves within the worlds of NLP.

Making sense of conversation: 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) [37], there's a good chance you've been left with a peculiar and distinctive experience. One that leaves you both impressed with how well the agent fares 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 she had not heard of the band before so we asked her to look them up. Pictured in this Figure 1 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—coreference resolution, recall of facts from a long, multi-turn conversation, is an unsolved and difficult problem in NLP [59]. However, things devolve quickly after responding to our typo-ed follow-up, asking "what did you (sic) thing?" Zo responds with one word, "Choctaw". Members of *A Tribe Called Red* descend from Cayuga and Ojibway peoples in Eastern North America, now known as Ontario, Canada [93]. What might make a bot believe that Choctaw, a people from Southeastern North America, is a reasonable response to such a question?

If an agent sees language as a set of symbols that have categorial or statistical associations, a bot might determine that the conversation in Figure 1 makes references closely associated with indigenous nations. Following that association, the bot responds with any indigenous tribe name it can retrieve [90]. Even if this is a reasonable turn from the chatbot's point of view, it is *not* a reasonable response from our side of the conversation. In fact, this response is problematic. It's disrespectful. The bot engages in race-talk, reproducing discriminatory, racist speech [5,14]. How can we help Zo and other bots do better?

We know that chatbots don'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 differences between thousands of indigenous nations into a single name stored in memory. However, just because Zo exists in a silicon space without our contexts does not mean that our contexts suddenly disappear. The contexts of our worlds are still present, whether a chatbot understands that or not.

Focus on syntax: What role does theory play?

The 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 [21,74]. 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 NLP domain.

This matters precisely because pragmatics contains 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 foci in NLP, including a turn to semantic grammars in the 80s, NLP is indebted to a focus on syntax [53,74]. In more recent NLP trends towards probabilistic variants of formalized grammars [67], language is constructed through a focus on the ordering of words probabilistically—an orientation towards language that is close to syntactic structuring. Ultimately, an abundance of racist content, and race-talk generally, is syntactically valid.

Without a growing technical capacity for context, we are not able to contend with the consequences of context at a structural level. We have been deferring the 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. A focus on word ordering does not remove context, even if context is unaccounted for. Zo need not understand the "trash heap" of pragmatics to draw from and contribute to the way the world is entangled in language [29]. Even without accounting for context programmatically, Zo is already acting from a position of agency and context, with its own machine intelligence [80].

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

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. If we want to mitigate some of this difficulty, we must modify our orientation towards human-machine conversation. Given the state of the art, how might we account for context? We need to consider how different types of actors, like humans and bots, with very different capacities come together to constitute talk that is collectively meaningful. If we take a machine's context and agency as a starting point, how can they contend with race in language?

It appears that an underlying assumption of a generalized chatbot like Zo is that bots can have conversations embedded in "universal" cultural contexts. But meaning and context are not universal. These constructs come to make sense through their specific and varied networked relationships. Even if we were to hold a generalized chatbot up to a human standard, what human can have a conversation on any topic, in every context, with anyone? These underlying assumptions are at odds with the issue of semantics and pragmatics. Excepting idle pleasantries, and asking for the time or the weather, it's unclear what "generalized chit chat" is.

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. Moreover, there is no reason that the number of agents in a conversation should be limited to one generalized, all-purpose chatbot. 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. Although this heterogeneous version of chatbot design might appear simple—certainly there exist many domain specific chatbots—consider how this idea expands as bots develop networked relationships through an ensemble. There is a world of possibility for what corresponding interactions might look like.

Consider a conversation where the bot you are chatting with —perhaps the conversation controller bot—realizes the talk may be slipping outside of their domain. 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. Thus, 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 rather than failures. Shortcomings might be presented through meta-data and reason-logs, citing issues like confusing language use or out-of-domain references. Confronted with an out-of-bounds topic, which might include race-talk, the bot could present a report with clarifying information rather than printing a statement of indifference that reproduces racist language. 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?

From their internal point of reference, bots learn to converse based on predictions that are consistent with their internal context rather than the world at large. 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 flippant contemplation, like turtles as a non-sequitur. If chatbots are to be more responsive to and responsible for inferences like this, it's clear that we need better ways of reconciling the differences between machine-internal and machine-external contexts, the context of the algorithm and the real-world outside. 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 at the cost of being able to understand or adjust their internal states. While prediction accuracy is enticing, an algorithm's internal conditions are critical in accounting for what is learned and how this learning becomes action. Understanding a bot's agency hinges on making sense of internal conditions. If we are unable to comprehend the agency of a chatbot, how will we build a deeper understanding of the differences between chatbot and human worlds, and how will we make the differences generative? To determine how we can leverage ML to handle race-talk better, we need to consider how human and machine agency impact a bots conversational learning.

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 using 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. "[Learning] must be done with the data that is available, not the data one would want" [18].

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 the creation of key guiding principles, taking a nod from previous U.S. policy setting [1,30,70,81]. A lot of this dialogue centers 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 between transparency—being able to see what is happening within a system—and making a system accountable [7]. 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, responsible algorithms?

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

Because machine learning is embedded in the language of abstraction, it can be difficult to make sense of how algorithmic processes connect back to our experiences of the world and to problem spaces like race-talk. While the following example is outside the problem space of race-talk directly, it concretely illustrates how the inner-contexts of algorithms are agentially contributing to machine-external contexts. Starting in the 1990s, ML algorithms have been studied and deployed for predicting the risk of pneumonia in a healthcare context [18,25,26]. 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 [18]. These studies compare the outcomes of a number of ML models, including a rule-based model and a neural net model. Unsurprisingly, the neural net was the most accurate model. But, it was ultimately deemed too dangerous to use with actual patients. Accuracy is not necessarily the best measure for evaluating a ML 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 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 abstraction, there are specificities of the machine-external context that pose problems for ML 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 [44]. 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. 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 work practice [27]? How do we account for this type of bias when developing and deploying virtual healthcare bots [12]?

Outside of healthcare, machine-external entanglements with race have major implications for algorithmic 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 [47]. 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 [79]. 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 engage with the structures of race and race-talk.

Interpretability and tunability: How do we leverage algorithmic agency?

Questions of algorithmic accountability are especially difficult in the context of neural nets. Neural nets—while often lauded as magically accurate—pose serious problems for understanding machine agency. There are two issues at stake, algorithms that cannot be probed and algorithms that cannot be adjusted to remedy a dangerous output, be it racism through medication dosage or abusive language.

There is growing research on interpretability of neural networks, but interpretability is not a panacea [6,61]. While encouraging, much of this research shares underlying assumptions with transparency, a severely limited construct

[7]. We agree with Ananny and Crawford [7] that asking for transparency—or interpretability—is not enough. What happens when there are problems with a neural net we don't know about or have no way of adjusting?

We need neural nets that are tunable. Nets (and ensembles) that can be adjusted and responsible to their networked technosocial context. When thinking about how these types of models can be tunable, we need to examine the ways neural networks are already 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 [36], 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 [50]. While these systems can be tuned at the time of training, these adjustments have not received in-depth research scrutiny. With a more developed 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 responsible to the network of worlds they participate in.

Probing and adjustment allow for responsibility. We are already participating in finicky behaviors with neural nets to help them converge or produce "optimal" outputs. 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 [18]. While advances in ML have resulted in high-accuracy, interpretable, and adjustable models that are a good-fit for healthcare datasets, these models do not fit text-based datasets well—the kinds that might be used for AI chatbots. Neural networks aren't going anywhere, we must learn to ethically and response-ably work with them if we are going to create chatbots more capable of handling race-talk.

Interdisciplinary partnerships: How can we champion equitable cross-domain collaborations?

There is an urgency to study algorithms that are already in use [15,22] and to study the entire development cycle for generating deep learning algorithms. If we take seriously the challenges of tunability, 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 field's 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 into the power of non-normative interpretation by providing a relational perspective on the agency of learning algorithms [58]. When knowledge in ML is valued equitably to knowledge in other fields, these research partnerships will allow us to develop more culturally responsive and responsible systems. If we embrace more intellectual inter-cultural exchanges with other domains, we open up the possibility of building algorithms that are more deeply connected to our varying technosocial ecosystems.

Returning to the question *how does a chatbot's agency impact its learning*, we are confronted with another essential question: *how do we best understand a chatbot's agency?* What steps should we take to perform basic research into responsible deep learning? Through interdisciplinary research ecosystems, we are more able to address the concerns of algorithmic accountability and build futures that value the plurality of contexts that exist. These long-form collaborations are vital for understanding and developing 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. In taking on Haraway's call to "stay with the trouble" we are holding onto the complexities of the worlds we have cut through. Critical issues cannot be addressed through neat separations between what people do and how machines operate. In determining where we go from here, we have to hold onto the complexities of our lived experiences, refusing to reduce the world into something that is uniform or singular.

Our acknowledgement of the trouble is a recognition that there is no *outside* of race. We are all bound up in the good and bad of life—not addressing the trouble does not make it go away. By talking about race and bots, we are working to make possible interactions that are more equitable and bearable. This requires us to stay with the *trouble,* widening the ways of living with humans, bots, and others [43].

Building Better Worlds: What is good enough?

Good enough will always be a moving target. As there is no one way to "solve" racism, we have tried not to be prescriptive about what steps should be taken. Trying to enable chatbots to better handle the complexities of race-talk is no small task and there is no silver bullet. Although there is no *one* good enough, there are *many* steps we can take in the development of chatbots that do better.

Steps like developing diverse databases, exploring ensembles of chatbots, and engaging in interdisciplinary collaborations give us a place to start. Rather than thinking about chatbots performing better or worse than humans, the goal is to develop bots that are capable of recognizing and responding to race-talk in the near future. The goal is not to solve an AI-complete problem, but to develop working solutions that can be achieved in the short-term. Solutions that can continue to be refined and developed in the long-term pursuit of good enough. As a starting point, good enough requires that we are actively engaging with how race and bias can manifest in our chatbots.

Whether or not race has actively been accounted for, artificial agents are already implicated in the structures of identity and race. Chatbots are being employed in research and in industry, often with the intention of building better worlds. There have been a range of studies presented at CHI that detail artificial agents involved in nursing, educational settings, activism, and conflict resolution [62,73,75,88,89]. These domains are deeply entangled with structures of race. While handling race-talk should not dwell on the human-like, a notable portion of this work is focused on if agents can achieve human-like abilities through talk and embodied presence. Putting aside the target of replicating human capacities—thoroughly debated and contested in ongoing conversations around the Turing Test [39,82]—our concerns center on the ways that technological artifacts, like bots, have politics [87]. Chatbots have their own mechanisms and agency. The more we focus on the algorithmic and agential potential chatbots already have, the more we will be able to start developing chatbots that handle race-talk more responsibly.

REFLEXIVE DISCLOSURE: RECOGNIZING OUR ROLE

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. We come from the United States and the United Kingdom, with backgrounds in sociology, psychology, HCI, computer science, digital humanities, science and technology studies, and critical theory. Like so many, our ethnicities are woven from a complex set of threads. We profit from the wealth and dominance afforded by living in and being educated in the West. Yet, at the same time we are composed of the legacies of colonialism and the subjugations of indigenous peoples. To ensure our voices are just part of a much larger dialog happening in this space, we have made 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.

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.

CONCLUSION: HOW DO WE EMBRACE THE TROUBLE?

In writing this paper, we set two essential questions to guide this work: 1) How can chatbots handle race in new and improved ways? and 2) Why is race-talk so difficult for chatbots? These questions have taken us down many paths to understand how race-talk is interwoven with technical configurations supporting chatbots.

An important contribution of this research is helping HCI practitioners understand how specific technosocial configurations are fundamentally entangled with their work. By drawing together technosocial interactions involved in race-talk and hate speech relative to databases, NLP, and ML, we strive to support the development of generative technosocial solutions——like a multiplicity of chatbots that upend the all-knowing agent. Chatbots are already exacerbating social harm specific to race. In working to mitigate these harms, there is potential for novel race-focused developments for chatbots specifically and for AI generally, like building off of work in raciolinguistics.

This work also makes contributions for HCI practitioners broadly concerned with identity, race, or equity in design. We demonstrate how social and technical conditions develop together in ways that must be reckoned with when forming human-machine interactions. For NLP and ML practitioners (and others who work with bots), seeing the connection between known problems and ethically critical topics like race is important. Hard problems in AI require practitioners to develop context specific solutions (i.e., focused on humor, language, or race). Staying with the trouble is not about neat resolutions. It is about embracing the complexities of our lives to enable better, though still troubled, paths forward. Clarifying a context, like race, and its manifestations can help guide these efforts.

Through making tangible the abstract and disparate qualities of race and chatbots, this paper works as a synthetic guide, pointing us towards possible futures where chatbots are more capable of handling race-talk in its many forms. The one question left is, *what is the racial context of your chatbot?*

ACKNOWLEDGMENTS

Thanks to our friends, reviewers, and colleagues who supported this work with their invaluable time, feedback, and challenges. And, a special thanks to everyone in the HXD group at Microsoft Research. Further, we are all indebted the many scholars and educators who have paved the path for this type of critical, reflexive research on race, identity, and justice. This research was partially supported by the NSF under Grant No. DGE-1148903.

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. H. Samy Alim, John R. Rickford, and Arnetha F. Ball (eds.). 2016. *Raciolinguistics: How Language Shapes Our Ideas About Race*. Oxford University Press.
6. David Alvarez-melis and Tommi S Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *EMNLP 2017*.
7. 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
8. 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
9. James Baldwin. 2010. As Much Truth as One Can Bear. In *The Cross of Redemption: Uncollected Writings*, Randall Kenan (ed.). Pantheon Books, New York.
10. Karen Barad. 2007. *Meeting the Universe Halfway*. Duke University Press, Durham.
11. 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
12. 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
13. Marcia Biederman. 2003. At $10 a Year, Automated Buddy Loses Laughs. *New York Times*.
14. Eduardo Bonilla-Silva. 2002. The Linguistics of Color Blind Racism: How to Talk Nasty about Blacks without Sounding “Racist.” *Critical Sociology* 28, 1–2: 41–64. http://doi.org/10.1177/08969205020280010501
15. 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
16. 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/
17. 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
18. 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
19. 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
20. 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
21. Noam Chomsky. 2002. *Syntactic Structures*. Mouton de Gruyter, Berlin.
22. Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*: 1–14. http://doi.org/10.1177/2053951717718855
23. 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
24. Beth Coleman. 2009. Race as Technology. *Camera Obscura* 24, 1.
25. 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
26. Gregory F Cooper, Constantin F Aliferis, Richard Ambrosino, and John Aronis. 1997. An Evaluation of

Machine-Learning Methods for Predicting Pneumonia Mortality.

27. 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

28. Kimberle Crenshaw. 1991. Mapping the Margins: Intersetionality, Identity Politics, and Violence Against Women of Color. *Stanford Law Review* 43, 6: 1241–1299.

29. Gilles Deleuze and Felix Guattari. 1987. *A Thousand Plateaus: Capitalism and Schizophrenia*. University of Minessota Press, Minneapolis.

30. Nicholas Diakopoulos. 2014. *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*. Columbia University Academic Commons. http://doi.org/10.7916/D8ZK5TW2

31. 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

32. Paul Dourish. 2014. No SQL: The Shifting Materialities of Database Technology. *Computational Culture*, 4: 1–37.

33. Jacob Eisenstein. 2013. What to do about bad language on the internet. *Naacl-Hlt*, Association for Computational Linguistics, 359–369.

34. 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

35. 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

36. 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.

37. 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/

38. 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

39. Barbara J Grosz. 2012. What Question Would Turing Pose Today? *AI Magazine* 33, 4: 73–81. http://doi.org/10.1609/aimag.v33i4.2441

40. 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

41. 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.

42. Donna J. Haraway. 1991. *Simians, Cyborgs, and Women: The Reinvention of Nature*. Routledge, New York. http://doi.org/10.2307/2076334

43. Donna J. Haraway. 2016. *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press, Durham.

44. 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

45. bell hooks. 2003. Talking Race and Racism. In *Teaching Community: A Pedagogy of HOpe*. Routledge, New York, NY, 25–40.

46. 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/

47. David Ingold and Spencer Soper. 2016. Amazon Doesn't Consider the Race of Its Customers. Should it? *Bloomberg*.

48. 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

49. 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

50. Natasha Jaques, Shixiang Gu, Richard E Turner, and Douglas Eck. 2017. Tuning Recurrent Neural Networks with Reinforcement Learning. *ICLR Workshop*.

51. 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

52. Daniel Jurafsky and James Martin. 2017. Dialog Systems and Chatbots. In *Speech and Language Processing.*

53. 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

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

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

56. 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

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

58. 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

59. 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.

60. 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

61. 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.

62. 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

63. Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3 LDC99T42. *Philadelphia: Linguistic Data Consortium*. Retrieved from https://catalog.ldc.upenn.edu/ldc99t42

64. Mark C. Marino. 2006. I, Chatbot: The Gender and Race Performativity of Conversational Agents.

65. Mark C. Marino. 2014. The Racial Formation of Chatbots. *CLCWeb: Comparative Literature and Culture* 16, 5. http://doi.org/10.7771/1481-4374.2560

66. 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

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

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

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

70. 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

71. 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/

72. 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/

73. 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

74. 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

75. 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

76. 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

77. 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

78. Caroline Sinders. 2016. Microsoft's Tay is an Example of Bad Design. *Medium*. Retrieved August 27, 2017 from https://medium.com/@carolinesinders/microsoft-s-tay-is-an-example-of-bad-design-d4e65bb2569f

79. Latanya Sweeney. Discrimination in Online Ad Delivery.

80. 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

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

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

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

84. Joseph Weizenbaum. 1966. ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM* 9, 1: 36–45. http://doi.org/10.1145/365153.365168

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

86. Keving A. Whitehead and Gene H. Lerner. 2009. When are persons "white"?: on some practical asymmetries of racial reference in talk-in- interaction. *Discourse & Society* 20, 5: 613–641. http://doi.org/10.1177/0306312706069437

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

88. 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

89. 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

90. Zhao Yan, Nan Duan, Jun-Wei Bao, et al. 2016. DocChat: An Information Retrieval Approach for Chatbot Engines Using Unstructured Documents. *ACL (1)*.

91. Microsoft Bot Framework. *Microsoft*. Retrieved August 10, 2017 from https://dev.botframework.com/

92. IBM Watson. *IBM*. Retrieved August 10, 2017 from https://www.ibm.com/watson/

93. A Tribe Called Red. *Tribal Spirit Music*. Retrieved from https://tribalspiritmusic.com/artists/a-tribe-called-red/

94. 2016. Tay AI. *Know Your Meme*. Retrieved June 1, 2017 from https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.0000fpjmog51cfpxpwz11olji2ndk