RESPONSIBILITY IN THE AGE OF IRRESPONSIBLE SPEECH

Benjamin Mitchell, William Fleischman

Villanova University (USA)

benjamin.r.mitchell@villanova.edu; william.fleischman@villanova.edu

ABSTRACT

We discuss the impact of language on some ethical problems surrounding the interactions of technology and society. We focus on problems of careless and irresponsible speech in the contexts of artificial intelligence and social media. As these areas are central to modern public discourse, the inappropriate use of language by computer professionals in these contexts has the potential for serious harm.

KEYWORDS: responsibility, language, artificial intelligence, machine learning, social media.

1. INTRODUCTION

The language used to express a concept is important, particularly when the concept is a new one. This is by no means a novel insight: Joseph Weizenbaum commented on this in the context of computing many years ago (Weizenbaum, 1972). In spite of this history, current public discourse suggests that a reminder and an update may be needed. From a linguistic standpoint, when a new concept is encountered, there are essentially two options; either an entirely new term can be created, or an existing term can be re-purposed. The latter is easier and more common, but when we attempt to re-purpose an already existing word into a new context, there is always some conceptual "bleed-through." Connotations of the original usage are ascribed to the new word even when they are not truly warranted.

Creating new words from scratch is difficult, and many attempts to get such terms into circulation fail. It is therefore perhaps unsurprising that the field of computing has a long history of simply borrowing conceptually linked terms and re-purposing them, rather than attempting to define new words. Even the term "computer" originally referred to a human who performed mathematical calculations as a career. But as convenient as repurposing existing terms is, there are clear hazards to doing so, and it must be the responsibility of computing professionals to ensure that the terms we use are not mis-interpreted by those who lack the background to directly understand their intended use.

Sometimes this careless use of language is unintentional, but frequently it appears to be done with malice aforethought. The dominant mode of political rhetoric in our society, for example, seems to revolve around the idea that perception is more important than truth, and that carefully selected terminology can be used to appeal to people's baser instincts while still maintaining some façade of impartiality. To take one example from a story currently dominating the news in the United States, informed reporters, government officials, and casual observers all commonly repeat the mantra "We've seen the transcript [of the phone call between the

presidents of the U.S. and Ukraine that may result in articles of impeachment entered against the American president]…" In fact, there are very few individuals who have actually seen the full, accurate transcript of that call, because its potentially incendiary nature led to the "reconstructed transcript" being quickly locked down on a server in the White House's most classified computer system. But by now, everyone in the public "understands" that the transcript of that call is a matter of common knowledge. This imprecision is convenient for those who wish to dismiss the importance of the conversation since at least one national security individual, who listened in on the call as a matter of his official duties, has openly criticized the omission of crucial words and phrases in the publicly disclosed "transcript." (Barnes, Fandos, & Hakim, 2019)

Whether inadvertent, reflexive, or calculated, imprecise or careless speech can have serious consequences and influence the thoughts and actions of individuals and collectives. In this paper, we consider the dangers of such speech in two distinct contexts: First, in public understanding of the capabilities and limitations of machine learning and, more generally, artificial intelligence; and second, in the ways in which careless and irresponsible speech by prominent executives of social media companies can undercut responsible behavior by computing and information professionals and frustrate efforts to find sensible measures to regulate the practices of social media platforms.

2. ANTHROPOMORPHIZATION AND SILICOMORPHIZATION

There is a complex interplay between science, science fiction, and public perception. Artificial Intelligence (AI) has always been deeply entangled in science fictional narratives. This manifests in many ways and affects both researchers and members of the public at large. The result is that many people's reasoning about the world is based on a mythologized version of AI that can lead to dangerous conclusions.

AI researchers have a long history of underestimating the difficulty of the field's core problems. In one memorable anecdote from the founding era of the field, several prominent computer scientists estimated that ***programming a computer system to replicate all the important functionality of a human mind might take several graduate students as much as a few months to accomplish.*** Some seventy years later, we have yet to even come close. This has not stopped popular portrayals of AI from ascribing human-like behaviors and capabilities to these systems. HAL 9000 from *2001: A Space Odyssey*, Skynet from *Terminator* and the eponymous cute robot from WALL-E are just a few of the many iconic examples. Whether implacable foe or compassionate helper, these systems are presented as being "not so different from you and me." Perhaps they have certain affective deficiencies (e.g. Skynet displays a total lack of empathy for the suffering of others), but nothing outside the range of behaviors displayed by actual humans (e.g. psychopaths display a total lack of empathy for the suffering of others).

These narratives paint a highly misleading picture of the capabilities of real-world AI systems. In actuality, all "Artificial Intelligence" systems to date are just purpose-built software tools designed to automate specific and narrowly defined processes. An "AI" has less in common with a human, and more in common with a kitchen knife; both are useful tools for assisting a human to get something done faster and better, but neither has any "agency" of its own. We give human names to these systems (Eliza, Siri, Alexa, Watson, etc.), although they are no more 'human' than a toaster. We use words that imply human-like thought processes (attention, understanding, belief, etc.) as labels for simple mathematical equations and algorithms that have only loose conceptual ties to the conventional meaning of the terms. Once these

anthropomorphic characterizations of "AI" are internalized by the public, sweeping extrapolations of these tools' potential are almost inevitable, resulting in misplaced trust in the capabilities of such systems.

The flip side of this exaggerated conception of the power of "AI" is something that perhaps deserves the name "silicomorphization," the reductive view of human intelligence and decision making based on the conflation of human intelligence with the operation of a digital computer. Naturally, in any comparison of capabilities based on this view, humanity comes off rather badly; a computer will always do a better job of being a computer. The result is a systematic denigration of the reach and richness of human intelligence and the robustness of human judgment. This devaluation of human capacity seems particularly harmful when taken as received wisdom concerning the relations between humans and machines, and the future of humanity itself. It is also a pillar of the myth of technological inevitability, which in many spheres serves to anaesthetize the conscience of those whose work involves the development and utilization of AI for purposes that are ultimately destructive of human values and human moral agency.

As we have indicated, AI provides useful tools for assisting decision making in cases where the context of the decision process is well understood and is seen to advance human well-being. But obeisance to technological inevitability in the form of unthinking substitution of automated decision-making for human judgment is fraught with danger. Once again, Joseph Weizenbaum understood the bargain: "Technological inevitability can thus be seen to be a mere element of a much larger syndrome. Science promised man power. But, as so often happens when people are seduced by promises of power, the price exacted in advance and all along the path, and the price actually paid, is servitude and [moral] impotence. Power is nothing if it is not the power to choose. Instrumental reason can make decisions, **but there is all the difference between deciding and choosing**." (emphasis added) (Weizenbaum, 1976)

3. IRRESPONSIBLE SPEECH IN THE CONTEXT OF SOCIAL MEDIA

One important consequence of the anthropomorphization of computer systems is to make it much easier to assign *blame* to these systems when failures occur. The purveyors of such systems often encourage this type of thinking, as it absolves them of culpability when harm is done, though they are quick to take credit when the results are good. Again, however, there is nothing truly "inevitable" about this line of reasoning, and there are plenty of reasons for pushing back against this narrative.

Helen Nissenbaum (1994) urges the adoption of a robust standard of accountability for computing professionals. She rightly observes that it is no more appropriate to assign blame to a computer system than it is to assign blame to any other tool; ultimately, accountability requires moral agency, and we should no more attribute moral agency to an algorithm than we would to any other technological tool. When a Boeing 737 MAX airplane falls from the sky, we might consider a variety of humans as worth investigating for possible responsibility (the pilot, the maintenance crew, the manufacturer, etc.), but we would never allow Boeing to place the blame on the plane itself. Yet when FaceBook's recommender system is found to be amplifying political disinformation, we are asked to believe that it is the *fault* of that algorithm, and FaceBook is merely a hapless bystander.

How can accountability survive in an atmosphere in which someone like Mark Zuckerberg can deny and distance himself and his company from one scandal after another? The examples are

numerous – the misappropriation of user data by Cambridge Analytica, the dissemination of false information in its newsfeed, and the strange policies regarding whether political advertisements may contain verifiably false information, to name just a few. In testimony before the U.S. Congress, when asked directly whether a political actor could run an ad containing politically inflammatory false statements, "Mr. Zuckerberg said the platform would take down posts from anyone, including politicians, that called for violence or tried to suppress voter participation." (BBC News, 2019) This, of course, is a reply to an entirely different question. Additionally, as far as its final clause is concerned, it is demonstrably false.

Among many other examples that can be adduced of evasion, shading the truth, or proclaiming ignorance of contentious action on the part of Facebook, we can cite Zuckerberg's narrow, legalistic denial that the exfiltration of personal information concerning up to 87 million users constituted a data breach, in flagrant contempt of the common understanding of the meaning of the term. And, when asked about Facebook's contract with Definers Public Affairs, a consulting firm it used in an attempt to discredit Facebook's critics, Zuckerberg claimed that "he did not know what Definers' activities were, or who at Facebook authorized that work. Probably 'someone on the communications team,' he offered." If ignorance, or the pretense of ignorance, is an effective defense for the well-placed, why should it not be equally available to the subordinate? (Lee, 2018)

Sadly, Facebook is merely the tip of the technological iceberg. YouTube has shown similar problems with irresponsible use of algorithms promoting falsehood and political bias (Lewis, 2018), and encouraging the radicalization of users (Ribeiro, 2019). In a response on their official blog, YouTube said "Our systems are…getting smarter about what types of videos should get [recommended less], and we'll be able to apply it to even more borderline videos moving forward." (YouTube, 2019) While this is perhaps better than simply ignoring the problem, it still amounts to an instruction to ignore past failures and blindly trust them to do what's best in the future. It is also worth note that this is presented as an improvement to an already great system; at no time do they acknowledge their responsibility for harm, or even that any harm has been done in the first place. Note also that in saying that the "systems are…getting smarter," the company is using anthropomorphic language to imply that the system itself has agency and responsibility here; the company is presented as merely an assistant who is helping the system, but ultimately cannot be held accountable for the system's actions. As noted by Tufecki (Tufecki, 2015), there is every reason to believe that these big tech firms are merely the most public, and therefore most studied, examples of a far more pervasive problem.

4. SOME GUIDANCE FOR THE PERPLEXED

The complexity of the technological systems is often used to justify a disconnect between stated intentions and observed outcomes. In fact, there are many recorded cases in which a reasonable and knowledgeable observer might agree that a certain outcome could be difficult to predict; emergent properties of complex systems are notorious precisely for their unpredictability. To a non-expert, nearly any technological system can be made to seem sufficiently complex that this defense has an air of plausible deniability. Since it is impossible to prove a negative, particularly where intent is concerned, we are left with the rather sticky task of evaluating when a denial is *sufficiently* plausible, and when it stretches credulity beyond our willingness to tolerate. It should come as no surprise that the result will differ from individual to individual, making consensus building on these issues problematic.

It is easy to say that this problem is complex, and that like all social issues it is not amenable to easy quick-fixes. But as with all such problems, this claim is disingenuous. Certainly, an ideal long-term solution would likely involve a combination of legislation, education, and an overall shift in cultural norms. Yet there are straightforward steps that can be taken in the near term to both begin to ameliorate the problem and to help create the conditions necessary for a more sweeping solution further down the line.

In spite of claims to the contrary, the problem of accountability is not one which has remained unsolved in other domains. The primary goal of engineering as a discipline is to take the chaotic world in which we live and create systems which will behave in understandable and predictable ways. A skyscraper is an extremely complex artifact, but while there may be some details that are difficult to predict (e.g. why that one room is always so cold), there are a wide range of behaviors we simply will not tolerate. If a skyscraper collapses, "we didn't expect (or intend for) that to happen" is simply not a sufficient defense. The simple fact that software is a newer technology than suspension bridges, airplanes, and skyscrapers does not mean that we need to tolerate a complete lack of accountability in software systems. It is possible to specify unacceptable behaviors (for which the developer might be held strictly liable) without requiring software that is perfect and 100% free of bugs.

Similarly, we cannot allow the blame to be placed on the systems themselves; a machine learning algorithm trained on a biased data set is no more "at fault" for its poor performance than a bridge constructed on ground too soft to support its weight is to blame for its own collapse. Responsibility requires moral agency, which only humans possess. In spite of their anthropomorphic name, "artificially intelligent agents" are tools, nothing more. Until and unless we develop true artificial general intelligence (an event which has been estimated as "about 20 years away" at every point during the last 70 years), ultimate responsibility for the behavior of any system must fall to the humans who design, create, test, and deploy that system.

In his prophetic paper, "On the Impact of the Computer on Society," Joseph Weizenbaum exhorts us, as computer professionals, to recognize that "[t]he nonprofessional has little choice but to make his attributions to computers on the basis of the propaganda emanating from the computer community and amplified by the press. The computer professional therefore has an enormously important responsibility to be modest in his claims." (Weizenbaum, 1972) In the context of modern AI and ML systems, it is particularly important to be humble not only in our explicit claims, but also in the claims implicit in our choice of language. By choosing humble language over hyperbolic, we can redirect responsibility to humans and improve the public understanding of the systems at the same time. As an example, perhaps YouTube could state that its systems are "being made better at maximizing their scoring function," or perhaps that "we are changing the objective to better reflect our desires". It might require a few extra words of explanation to replace the term "smarter," but it would lead to a much more useful discourse on the matter (particularly if it led the public to question *whose* desires are being reflected by that objective function).

Until the tech giants can be held to this standard, the mid-20th century judgment of Friedrich Dürrenmatt seems uncannily pertinent to our moment in history: "In the Punch-and-Judy show of our century … there are no more guilty and also, no responsible men. It is always, 'We couldn't help it' and 'We didn't really want that to happen.' And, indeed, things happen without anyone in particular being responsible for them. … That is our misfortune, but not our guilt… Comedy

alone is suitable for us." (Dürrenmatt, 1964) The comedy, alas, is often of a rather mordant nature (in which we are the bitten.)

REFERENCES

Barnes, J., Fandos, N. & Hakim, D. (2019, October 29). White House Ukraine expert sought to correct transcript of Trump call. *The New York Times*, Retrieved from https://www.nytimes.com/2019/10/29/us/politics/alexander-vindman-trump-ukraine.html

BBC News (2019, October 24), Facebook's Zuckerberg grilled over ad fact-checking policy, *BBC News*, Retrieved from https://www.bbc.com/news/technology-50152062

Dürrenmatt, F. (1964, at 31), *Problems of the Theatre*, translated by Gerhard Nellhaus. Grove Press, New York.

Lee, D. (2018, November 16), Mark Zuckerberg, missing in inaction, *BBC News*, Retrieved from https://www.bbc.com/news/technology-46231284

Lewis, P. (2018, February 2), 'Fiction is outperforming reality': how YouTube's algorithm distorts truth, *Guardian News*, Retrieved from https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth

Nissenbaum, H. (1994), Computing and accountability. *Communications of the ACM*, vol. 37, no. 1, pp. 72-80.

Ribeiro, M., Ottoni, R., West, R., Asmeida, V., & Meira, W. (2019), Auditing radicalization pathways on YouTube, *arXive*, Retrieved from https://arxiv.org/abs/1908.08313

Tufecki, Z. (2015), Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency, *Journal on Telecommunications & High Technology Law*, pp. 203-218.

Weizenbaum, J. (1972), On the impact of the computer on society: How does one insult a machine? *Science*, vol. 176, no. 4035, pp. 609-614.

Weizenbaum, J. (1976, at 259), *Computer Power and Human Reason*, W.H. Freeman, New York.

YouTube (2019, June 5), Our ongoing work to tackle hate, *YouTube Official Blog*, Retreived from https://youtube.googleblog.com/2019/06/our-ongoing-work-to-tackle-hate.html