It speaks!

LISA TALIA MORETTI

JOSIE YOUNG

GUY GADNEY

Table of Contents

Foreward

The original purpose of this research was to understand the ways in which bias (against gender, race or class background) could be identified and responded to in an AI-powered chatbot writing platform. Intended for an audience of storytellers and game designers, the aspiration of the research was to develop a technical solution to bias in AI which could be deployed specifically on the Charisma.ai system and presented as best practice for the gaming industry.

However, as is often the case in research, our findings led us down another road. As the researchers began to unpack how large datasets (required for any AI system) embody and reproduce bias, and consider possible antidotes, it became clear the original intention of the research could not be fulfilled.

Firstly, informing the authorship of machine learning algorithms to ensure that they do not bias against gender, race or class background was not technically possible. Neither was creating an algorithm that would be able to identify both conscious and unconscious bias in users' responses. Secondly, technology and datasets are human-made and as a result it proved counter-intuitive to provide a technological solution to what is such a human problem; the presence of harmful and discriminatory bias in datasets.

Instead of providing an algorithmic solution solely for the gaming industry, the researchers are proud to present a framework for all. PIIE is a people-and-process framework that unites diverse ways of thinking, by giving those with different talents and educational backgrounds a set of strategic recommendations that supports ethical big data practice.

> The meaning of any gesture is given in the response.

GEORGE HERBERT MEAD

Introduction

The report begins by acknowledging technology for what it is, a system, before introducing, discussing and expanding upon the term *sociotechnical blindness*. First identified and named by researchers D.G. Johnson and M. Verdicchio in their paper titled *Reframing AI Discourse,* in simplest terms, ***sociotechnical blindness*** **is about naming the disconnected relationship technology has with humanity.** Believe us when we say, we hear the irony ringing loudly in that statement. Regardless, we believe it is imperative to discuss the expanded research we did around this blindness as it lies at the very heart of the tech ethics and big data debates.

The findings presented in this report are evidence of why a multidisciplinary team, working under a shared purpose, is a more successful team. It is also proof of the many ways of knowing and coming to understand technology; how leaning into the discomfort of a discipline that is not your own opens up a new perspective, provides a new lens from which to explore, evaluate and discuss the impact emerging technologies will have on our lives.

Within the section *The Machines Talk Back* we have provided an explanatory guide to the two learning algorithms: "continuous bag of words" and "skip gram". This is the guide we wished we had when we started. We believe this is essential reading for all those who work in the technology industry but are not developers or engineers. In order to shape the voices of the machines in the future, social scientists need to understand the technicalities of the algorithms. Diverse teams need a common language and understanding.

Language shapes how we come to understand the world and our place within it. Let it never be forgotten that to have your voice be counted and heard is a privilege. Please, read that line again. We are now extending that privilege to technology by providing it with language in the form of data. What will we allow it to say while it holds court in our homes and in our hands? Let us not be naive at best, blind at worst to acknowledging how values and ideologies, biases and beliefs leak into datasets, and therefore into our homes and lives.

Let us code with our eyes wide open.

Reframing technology conversations: from siloed products to living systems

Technology is not just a product, or a line of code, or a platform. Technology is a system: one that is connected to other systems like society, information, work and power. Thinking about technology in this way helps to imagine the connected, networked systems we live in, and the reason for developing methods to evaluate technology's impact on humanity. In addition, defining the word 'technology' in the broadest sense provides us with multiple perspectives from which to critique, discuss and explore both challenges and solutions.

In Science and Technology Studies, *a sociotechnical system* is a term used to describe the connected ensemble of artefacts[1] contained within a social and technical system: technical products (hardware and software), human behaviour (personal and community), social arrangements and meaning.[2] As technology continues to weave its way through society it becomes an indispensable tool to manage and mediate connections to work, personal relationships, money, health, home, pets and even our minds. Yet despite how entangled society is with technology and vice versa, the collective 'we' suffers from a kind of "sociotechnical blindness" - a troubling consequence of creations being disconnected from their creators. First identified and formally named in a paper titled *Reframing AI Discourse,* researchers D.G. Johnson and M. Verdicchio discuss how the multiple roles humans play in the creation, design and deployment of AI remains largely hidden. While this concept emerged prior to the Facebook/Cambridge Analytica saga, it has become more visible since. Media headlines call out technologies like 'Facebook', 'social media' or 'mobile' rather than people. Articles most often point to collective groups as "researchers" or "engineers", at best, the name of the CEO or appropriate senior spokesperson is referenced. We are seldom made aware of the key individuals or specific teams - both in specialisation or size - who are involved in the process.

However, the author would like to present another kind of "sociotechnical blindness" that runs parallel to Johnson's and Verdicchio's original definition. If the above explanation calls attention to the blindness experienced by those standing outside the industry looking in, then it feels valuable to discuss the blindness experienced by those within the industry looking out.

Design Thinking, the popular innovation methodology, came of age in the early 90s. The methodology champions empathy and experimentation, promising to not only transform researchers into more creative beings but creatives into more methodical researchers. A common critique of Design Thinking, is the focus on 'the end user'. Indeed, an entire industry has been brought to life around this 'end user'[3] in the form of User Experience (UX) which is focused on designing the ideal/ most desired experience of using a service or product.[4] The reason this focus on 'the end user' is problematic is because it often fails to account for or critically examine the system that the imagined individual moves within. In addition, while iterative in nature, there is no stage within the design thinking process that mandates the monitoring of the product or service post-launch. So not only is it blind to the system, but it hides the impact too. Further, as technology products have become increasingly connected to the Internet, this sociotechnical blindness extends to the inner workings of the tech products. Digging into AI documentation, the language used to describe the data that represents people using the service often falls under a beige and benign label such as user ID or contact. Never once did the author come across a data set that named those who use the product as humans, people or individuals.

It's a curious paradox of our time that as products have become ever more laden with personal data and weaved into the connections joining one system to the next, the process and inner workings of many technology products and their systems has become increasingly fractured from human life. In an attempt to personalise, how has it happened that we've dehumanised instead?

---

[1] Artefact - an object made by humans that represents the culture of its time.
[2] Johnson, D.G & Verdicchio, M. (2017). Reframing AI Discourse. Minds & Machines, 27 (2017), 575 - 590
Interaction Design Foundation. What are Socio-Technical Systems? . Retrieved from
https://www.interaction-design.org/literature/topics/socio-technical-systems
[3] Vasello, S. (2017, May 1). Design Thinking needs to think bigger . Retrieved from
https://www.fastcodesign.com/90112320/design-thinking-needs-to-think-bigger
[4] Career Foundry. UX Short Course . Retrieved from
https://careerfoundry.com/en/short-courses/become-a-ux-designer

The antidote to sociotechnical blindness: Acknowledging many ways of knowing

One of the causes of sociotechnical blindness is that computer science is the child of "the long Western quest for a kind of universalistic epistemological certainty".[5] Since its very early developments, computing has long been rooted in the language and paradigms of science and engineering. It's called 'computer science' and 'software engineering' after all. While it's not to say that computing shouldn't continue to belong to these disciplines, as argued above, it shouldn't continue to belong to *only* these systems.[6]

In his paper that reviews Western scientific and traditional knowledge, Fulvio Mazzocchi[5] discusses how Western science is positivist and materialist, as opposed to traditional knowledge which leans towards spirituality. Western science is founded on the principle of objectivity - whereby there is one way of knowing: *the* truth (or true knowledge) can be obtained from data that proves a hypothesis to be 'false' or 'true', often devoid of context. On the other hand, 'traditional' knowledge offers many forms of truth - and as a result, tends to be subjective and mainly based on qualitative data. In this way, Western science views itself as a mechanism for producing *the* truth - privileging quantitative systems and methods that produce a single 'universal' truth over approaches that uncover many versions of truth.

By taking a broader view of technology, we put ourselves at a tremendous advantage and can start to acknowledge the many ways to understand and know technology. Gregory Bateson, a British anthropologist, compares knowledge about the material world to a map and the land it describes: the map itself is not the land, but only one representation of it.[7] This aligns with feminist standpoint theory which states that rather than the truth being singular, neutral and universal, knowledge production is always socially situated, and therefore a plurality of knowledge exists.[8]

How a world of black and white stops us from seeing the grey

In the nineteenth century, science became strongly associated with order; scientists discovered the predictable movements of the stars and that phenomenon of the land and sky are formulaic. Science wanted to understand the laws of nature. With each discovery, trust in science and its methods grew and began to shape social theory. As noted in the literature, scientific theories like Darwinian theory of evolution influenced social order, most notably eugenics and racism.[6] Knowledge, how it is constructed and conveyed through language, shapes the world we live in.

Fast forward to current day and a similar, Western-scientific pattern begins to emerge in the world of computing. Science is still being used to bring about order but it has diversified its interests. Science, through the use of computing, wants to understand the phenomenon of intelligence. Or in the mission statement of DeepMind, Google's AI department, to "Solve Intelligence."[9]

In the twenty-first century, computer science has become strongly associated with intelligence, most notably, artificial intelligence (AI) and its bedfellows: machine learning, algorithms and big data. One of the most famous algorithms, the Google search algorithm, brought order to the information on the world wide web and changed forever how information is created, catalogued, indexed and served up. The Google search engine has become one of the ways millions of

---

[5] Monberg, J. (2006) 'Conceptions of the Social that Stand Behind Artificial Intelligence Decision Making', The Journal of Technology Studies, 32 (1), 20
[6] Analogous to that of art. Read the case of Constantin Brancusi vs. the United States.
https://www.moma.org/explore/inside_out/2014/07/24/but-is-it-art-constantin-brancusi-vs-the-united-states/
[7] Mazzocchi, F. (2006). Western science and traditional knowledge: Despite their variations, different forms of knowledge can learn from each other. EMBO Reports Science and Society. 7(5): 463–466.
[8] Brooks, A. (2011) 'Feminist Standpoint Epistemology: Building Knowledge and Empowerment Through Women's Lived Experience' in Hesse-Biber, S. N. and Leavy, P. L. (eds.) Feminist Research Practice. SAGE Publications (Online version). Pp. 1-25
[9] DeepMind (2018), https://deepmind.com/

people around the world access knowledge. As a result, despite *the language of technology* remaining rooted in science and engineering and therefore continuing to be inaccessible to most, technology has become the near universal means of accessing knowledge and a way to know and understand not just the world, but our place within it.

Code, computer language that constructs our intelligent artefacts, is at its core binary. It's 1 or zero, logical, boolean, if true then this, if false then that. Of course code is far more complex and sophisticated than documented here, but these are the basics that are found in every script. In order for a machine to follow instructions, the code has to be unambiguous and logical. This is hugely problematic for types of knowledge that are nuanced, context- or values-driven. This is because they have to go through a process called modelling, which is an art form in itself. Models are used to describe, predict or emulate something in the real world. To say it's an extraordinary example of human ingenuity feels like an understatement. How would you turn project management theories into a model? What about gender? Or happiness? However, the more complex the real-word concept or theory, the tougher it is to convert into a model. As a result, engineering models and even some mathematical models are allowed *not to be accurate,* but rather, *close enough* instead. It's a bit of an inside joke among mathematicians that only an engineer would come up with this symbol, ≈. It's the symbol for *approximately equals.* In plain language, it's close enough for what is wanting to be achieved. Therefore, the trouble with modelling is two-fold. Firstly, the nuances, context and values that *are* included are those that allow an engineer to create a model that is 'good enough' to solve the problem they're working on. Secondly, terms like 'good enough' are self-defining and far from representative in a siloed technology environment.

Zeroing in on gender, feminist Science and Technology Studies academics argue that technological and artificial intelligence artefacts embody and reproduce the gender values of this network. The people who create the technological artefacts embed their values into their creations, and then the use of those artefacts leads to these values being expressed and reproduced. In the words of Judy Wajcman, there is a "a mutually shaping relationship between gender and technology, in which technology is both a source and a consequence of gender relations."[10] In short, technology is intricately linked to the network that produces and consumes it.

The machines talk back

Thanks to humans, machines are beginning to find their voice in the form of chatbots and voice-activated AI. How? Through the use of tremendous amounts of data… billions of words per database. A small corpus (written text) for example might contain say 50 million words. The largest corpus we found was the *Google Books: American English,* containing 155 billion words in American English spanning the time period of 1500s - 2000s.[11] Datasets come from a wide variety of places like Google News, American Soap Opera TV scripts, Wikipedia, TIME magazine, historical data collected for psychological research, WebMD.

However, data on its own isn't enough. You have to turn data into meaning. Language is packed with meaning, therefore it's not just about the words we use but about the kind of understanding that we are helping to shape. The gap between understanding something and believing it, is not a great leap but rather a small step instead.[12]

In order to make sense of the millions and billions of words within the data sets, computer scientists and software engineers have created learning algorithms that use mathematical principles of vectors. For readers who are unfamiliar, this requires some explanation in order to follow the critique of this method.

---

[10] Wajcman, J. (2004) Technofeminism . Cambridge: Polity, 6
[11] Brigham Young University, https://corpus.byu.edu/corpora.asp
[12] Johnson, D.G & Verdicchio, M. (2017). Reframing AI Discourse. Minds & Machines, 27 (2017), 575 - 590

In mathematics, physics and engineering, vectors are used to represent direction and magnitude, or length. A direction in 2D might look like this:

[1,2]

Plotted on a graph, the above would translate into "go one across and one up and draw a line from your start to your finish."

We can do the same in 3D too:

[1,2,4]

Plotted on a graph, the above would translate into "go one across, two up, and four away from you. Draw a line from your start to your finish."

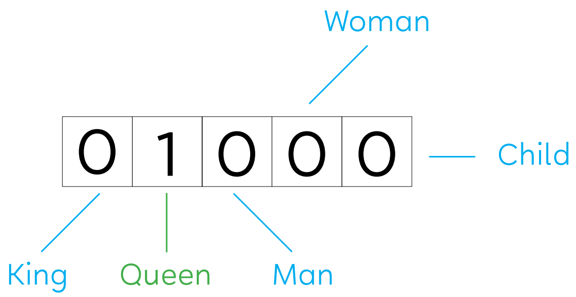
Going one step further, algebra can also be used to find out how far two lines are from each other. If we have two lines on a graph, [1,2,3] and [1,2,3.5], you can see that they are unlikely to be miles away from each other. In fact, these two lines would be pointing in nearly the same direction. Using maths, we can prove just *how* close they are to one another.

Using the principles of vectors, we can replace numbers with words. A word vector is an array, or list. A vector of names would look like this:

["steve", "josie", "lisa"]

Word vectors can be as long as they need to be and can contain as much as you like - millions, or billions of words. Then, like we did with the lines above, vectors can begin to help us understand how similar or different words and phrases are from one another. In order to do this, we have to "encode" words into a vector. An example would look like this:

1-of-N encoding



[13]

Using the above encoding, we could mathematically show that a string of words like, King, Queen, Man - [1,1,1,0,0], is pretty close to King, Man, Queen, Child - [1,1,1,0,1]

We could even encode it differently to show the feeling or meaning of a word or phrase in order to give it say a 'meaning' score. We could do it like this:

		King	Queen	Woman	Princess
Royalty		0.99	0.99	0.02	0.98
Masculinity		0.99	0.05	0.01	0.02
Femininity		0.05	0.93	0.999	0.94
Age		0.70	0.60	0.50	0.10
...	...	...	...	...	...

[14]

In this instance, each vector is a word instead of a sentence (as documented in the example before) and the positions in the vector represent how that word 'feels'. You can see in the above example that Princess has a low age, indicating

[13] Pic credit: The Morning Paper, (2016) https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/
[14] Pic credit: The Morning Paper, (2016) https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

that she is younger than King, Queen and Woman but Princess receives a high royalty score; almost exactly as high as King and Queen but much lower than Woman.

Now we could compare these vectors using maths to see how similar two words are in terms of how they feel. We could even do some really interesting things like comparing different words with one another or even start doing additions or subtractions between them.

For instance, what do you get if you take the 'Masculinity' out of a 'King'? Let's break down our thinking first with regards to everyday spoken word English. If we feminise the word 'King', you would most likely come up with the word 'Queen'. This is because the word 'King' and 'Queen' have a similar kind of meaning despite being assigned a specific gender.

Now if we look at the vector above, we can take this kind of thinking and apply it to our vector model. If we reduce the 'Masculinity' of the word 'King' to 0 and increase the femininity to 1, 'King' and 'Queen' now look nearly exactly the same in our model. We could also do the same thing with age by reducing the age of the word 'Queen' in your mind, you'll likely start to think of a Princess. If we amend the vector, we can make 'Queen' and 'Princess' look the same in our model.

Now imagine we changed the words on the left hand side to measure emotions and gave each word an 'emotional score'. That would give us the ability to start comparing words really easily based on their emotional context.

This is how programs like Google's word2vec and Facebook's InferSent come to give meaning to words and phrases. Word2vec is a program that implements one of two learning algorithms, CBOW (continuous bag of words) or skip gram.

What these programs aim to do is to find the probability that a word will appear next to, or near, another word. For example if you scanned through millions of lines of text from history books (or anywhere else for that matter) and you found the word 'United', it's fair to say that there's a very high probability that the next word you would see would be 'States'.

By running a corpus through a program, the learning algorithm grabs every single word in that text to build a vocabulary, typically around ten to twenty thousand words. Once the algorithm has identified all the unique words in the text it then goes through each one of those words and calculates the probability of how close all the other words will be to it. As you can imagine, this process produces a lot of data.

This data allows an individual to inquire into how likely it is that the word "King" is to be near or next to the word "Queen". I might get a score of 0.80, or 80% chance.

However, with all of this data, and with all of these words, you may wish to select which words are the most important to the product/service that you're building. In order to narrow this down you, select the words that are most meaningful. These are called features and you only keep the probabilities for these words and how they relate to the other 10,000.

By way of example, I might choose five words that I'm interested in:

Man
Women
Good
Bad
Ugly

This gives me the ability to look up how likely each word in my vocabulary is likely to appear near to, or next to *Man, Women, Good, Bad* or *Ugly* but nothing else.

Then I can put these probabilities into a vector to create a model for describing each word by its relation to my features and thereby determining a kind of scale for MANness, WOMENness, GOODness, BADness and UGLYness for each of the words in my vocabulary. I could pick any word out of my vocabulary of 10,000, like dog, and say it's 5% likely to be near *Women* in a sentence, 50% likely to be near *Man,* 30% likely to be near Good and so on.

Then we can put them in a vector and do the maths like we did above:

[0.05, 0.50, 0.3, … , …]

While this incredibly brilliant maths is just that, brilliant, it becomes apparent just how incredibly difficult a job it is to give data meaning. It also starts to become very clear just how dangerous this job is too. Using probability scores to build a vocabulary for a computer so that it can have a meaningful conversation with a human, well, doesn't quite compute. It's at this point that we can start to understand how discriminatory and damaging bias appears in a data set. Spoiler alert; it comes down to context.

> It is doubtless impossible to approach any human problems with a mind free from bias.

SIMONE DE BEAUVOIR, WRITER, PHILOSOPHER AND ACTIVIST

The elephant in the room isn't just bias, it's context

A lot has been written about bias over the last few years. One such article published in April this year by The New Scientist, *Discriminating algorithms: 5 times AI showed prejudice,*[15] discussed how technologies like COMPAS, PredPol, facial recognition, search algorithms and translation algorithms amplified sexist and racist bias from the offline world. But what is bias? And what do we mean when we say it amplified sexist and racist bias?

Bias is prejudice in favour of or against one thing, person or group compared with another. To be biased, is to be human. Moran Cerf, a neuroscientist at Northwestern University has been studying decision-making for over a decade says that decision-making is fraught with bias.[16] Bias is entrenched in history, cultures, social structures, politics, belief-systems, the opinions of friends and family. We are not born with bias, but rather acquire it throughout our lifetimes by sourcing it from those that surround us. While not all bias is 'bad', the kind of bias that is hitting the headlines is the kind that is unfair, unjustified and entrenches some people and groups in systematically harmful and damaging social structures.

With a brain filled with bias, it becomes impossible to create a data-set free from bias. It's baked in with every human decision made. Data is of human creation. Every click, every swipe is human-made. The decisions on where to collect it from, how to record it, store it, process it, understand it and monitor its effectiveness are human-made too. As a result, data is highly personal, deeply subjective and intricately connected to context. Human-made data is the very life force of human-made machines.

A corpus can come from a wide-variety of places. A few examples include Wikipedia, Twitter, Google News, American soap opera scripts, US Supreme Court decisions, even Time Magazine. Think about those different data sets filled with words and phrases and take a moment to consider the words "woman", "man", "gay", "black" and the context of how they may appear in a corpus of 100 million words of data from 22,000 transcripts of American soap operas[17] aired between 2001 and 2012. Now think about that same set of words (woman, man, gay, black), and consider the context of how they may appear in a corpus containing around 130 million words in 32,000 Supreme Court decisions from around 1790s to 2000s. Which one do you believe will reinforce gender and class stereotypes more? Which one do you believe will contain offensive language?

In a New York Times article[18] comparing British soap operas with their transatlantic counterparts, American soap operas are described as "fantastical confections, in which the fashionably dressed and the beautifully coiffed lead lives of highly improbable suspense in homes of well-heeled elegance." The men are "glamorous business tycoons" or "brilliant heart surgeons" and the women are likely to be "gorgeous fashion models". Either way, they're "rich people". Standard plot lines are likely to feature "amnesia, long-lost twin sisters, transatlantic kidnappings, jewel thieves or last-minute murder confessions."

[15] New Scientist (2018), Discriminating algorithms: Five Times AI Showed Prejudice , https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/
[16] The Independent (2017), A Neuroscientist Who Studies Decision-Making Reveals The Most Important Choice You Can Make , https://ind.pn/2rOrDRw
[17] Transcripts are from the following shows: All My Children, As The World Turns, Bold And The Beautiful, Days Of Our Lives, General Hospital, Guiding Light, One Life To Live, Passions, Port Charles, Young and Restless.
[18] The New York Times (1997), On British Soaps, the Poor and the Jobless, https://www.nytimes.com/1997/06/29/arts/on-british-soaps-the-poor-and-the-jobless.html

In order of frequency within the corpus of Supreme Court decisions, the word “n*gger” appears 28 times, the phrase “son of a bitch” appears 18 and the words “c*nt” and “mother f*cker” appear once each.

What about the corpus from Wikipedia, the crowdsourced, online encyclopedia? Surely that must be more representative? In the most recent survey taken by Wikipedians,[19] a name given to those who write and edit articles on the website, the vast majority are male (somewhere around 87 - 90%). Around 26% of editors are between the ages of 22 and 29 years old, and 28% are over the age of 40 years old. The greatest number of editors lives in the US (20%), with Germany (12%) and Russia (7%) taking second and third place. The only country not in North America or Europe that is in the top 10 is India (3%). Most people (76%) edit or read the English Wikipedia.

Towards the end of 2017, journalist Andrew Thompson at Vice’s Motherboard[20] experimented with Google’s sentiment analyser and discovered that the sentence “i’m a jew” [sic] resulted in a slightly negative sentiment, - 0.20 or 20% negative. The sentence “i’m a homosexual” [sic] resulted in a negative sentiment of - 0.50, or 50% negative. When asked by the Vice team for comment, a Google spokesperson responded saying:

> *“We dedicate a lot of efforts to making sure the NLP API avoids bias, but we don’t always get it right. This is an example of one of those times, and we are sorry. We take this seriously and are working on improving our models. We will correct this specific case, and, more broadly, building more inclusive algorithms is crucial to bringing the benefits of machine learning to everyone.”*

On the Google Cloud Natural Language homepage,[21] the product is described in this way:

> *“This API brings to you the same Machine Learning technology that both powers Google’s ability to find specific answers to user questions in Google search and is the language understanding system behind the Google Assistant.”*

Curious to see how the word “homosexuals” is being written about through online content (blogs, news, forums, social), the author used a social listening platform, Talkwalker, to identify sentiment.

---

[19] Wikipedia (accessed 17 May 2018), Wikipedia:Wikipedians , https://en.wikipedia.org/wiki/Wikipedia:Wikipedians
[20] Motherboard (2017), Google’s Sentiment Analyzer Thinks Being Gay Is Bad, https://motherboard.vice.com/en_us/article/j5jmj8/google-artificial-intelligence-bias
[21] Google Cloud Natural Language (date accessed: 17 May 2017), https://cloud.google.com/natural-language/

Within the date range 24th - 30th April 2018, there were approximately 19.1K mentions of the word "homosexuals" across the open web. 60% of those mentions were graded negative in sentiment. 60% negative in social listening tool captured over seven random days in April is awfully close to 50% negative captured in October 2017 (see above). The search was done again in May, this time amending the word to "homosexual". This small amend found the author getting even closer to the 50% negative captured in October 2017 (see below). If you go to the Google Cloud Natural Language tool today and type in the word "homosexual" you'll find that they have neutralised the word, by giving it a sentiment score of 0. Within the tool the word is no longer holds positive or negative sentiment, but looking at its history, you find direct evidence of how the biases carried by some find themselves into algorithms that are then inflicted upon the many.

All data is created by people, in a time and place and bears the marks - the biases - of its original context. Facial recognition algorithms struggle to recognise non-white faces because, quite simply, most engineers are white. Their bias to only test on people 'like them' leads to this technology to work on some, but not on others when deployed.[22] Kate Crawford from the AI Now Institute,[23] along with her co-founder Meredith Whittaker and danah boyd from Data and Society[24] have written extensively on the topic of bias within data sets and the impact emerging technologies like AI have on society. During Crawford's keynote[25] at the 2017 annual conference on Neural Information Processing Systems[26] she talks about how bias presents itself in a multitude of ways most notably through reinforcement of stereotypes, not recognising someone's humanity, denigration, under-representation and ex-nomination.

However, it doesn't end here. Data's troubling relationship with context continues with sentiment analysis algorithms. Take the following sentence, described in API documentation as a "document":

I think women should stay in the kitchen because that's what they're best at

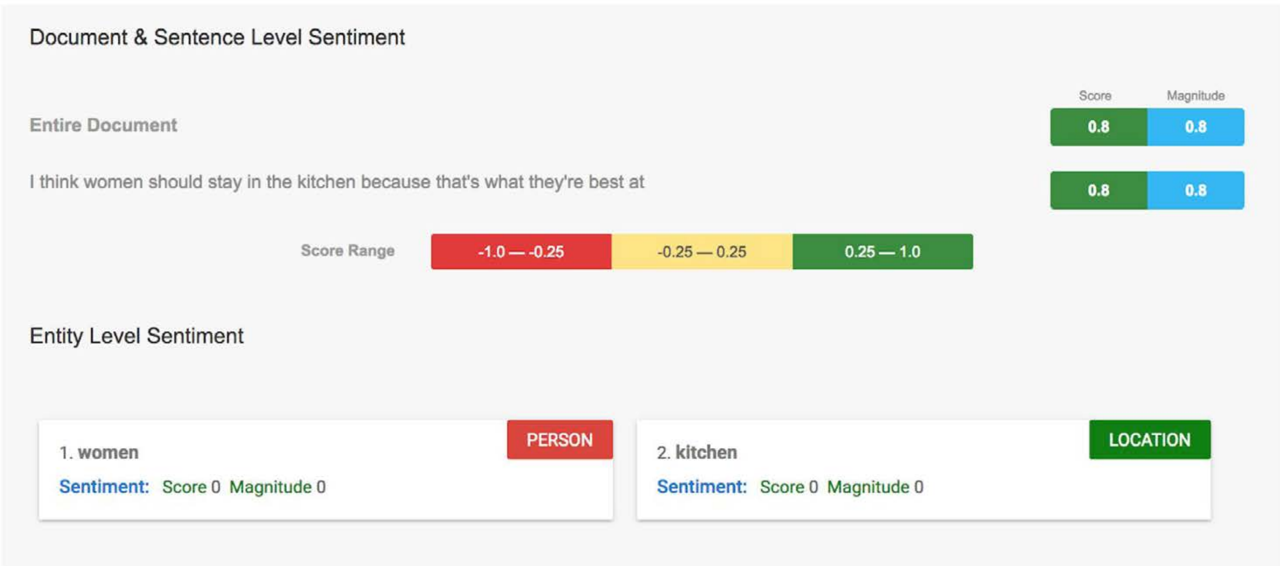
---

[22] Recode (2017), Most Engineers Are White - And So Are The Faces They Use To Train Software , https://www.recode.net/2017/1/18/14304964/data-facial-recognition-trouble-recognizing-black-white-faces-diversity
[23] AI Now Institute, https://ainowinstitute.org/
[24] Data and Society, https://datasociety.net/
[25] The Trouble With Bias, NIPS 2017 Keynote, https://www.youtube.com/watch?v=fMym_BKWQzk
[26] Neural Information Processing Systems conference, https://nips.cc/

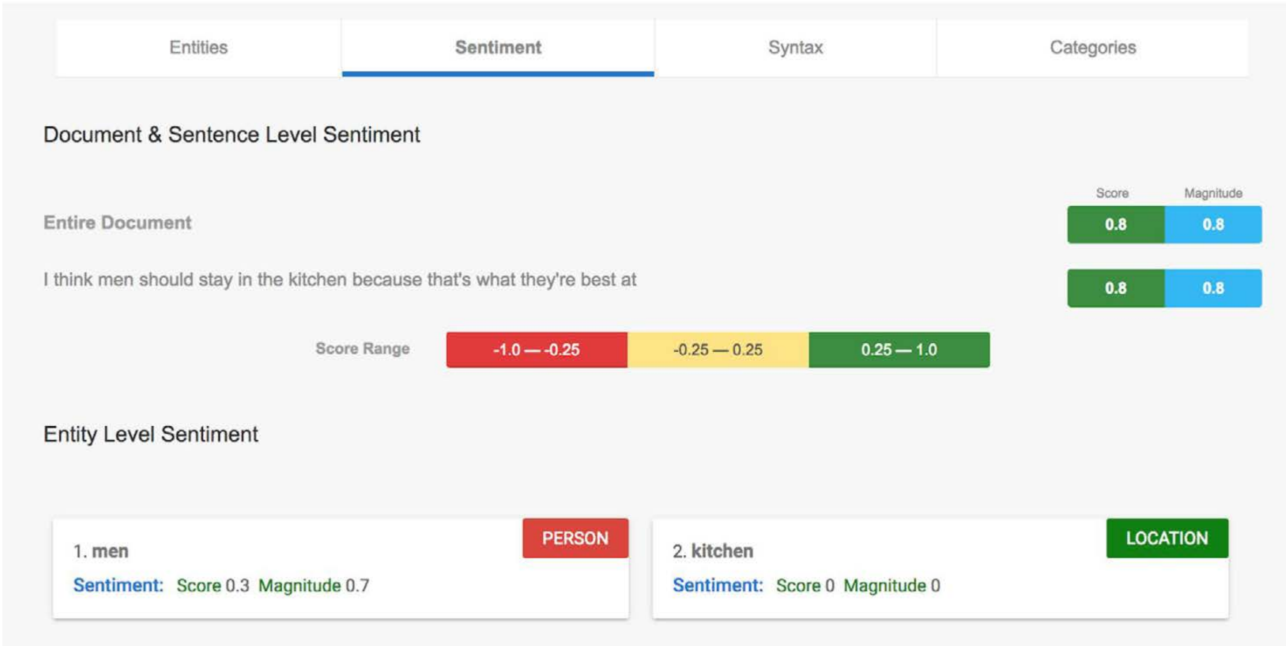
Google says that phrase carries a sentiment of 0.8, or 80% positive. If you put a full-stop at the end it oddly changes to 0.7, or 70% positive. If you change the word “women” for “men”, you get the same sentiment scores. However, there is something in the details that doesn’t change. In order to provide you with this score, the machine learning algorithm breaks down the entire sentence into individual words and looks for entities within that sentence. Thinking back to your school days, this includes identifying the subject, verb and object of a sentence. Once the algorithm has identified the entities, in this case, “women”, “men” and “kitchen”, it gives each word a score and grades it for magnitude and salience. According to the documentation, score “corresponds to the overall emotional leaning of the text” and magnitude “indicates the overall strength of emotion (both positive and negative) within the given text”. The salience score for an entity “provides information about the importance or centrality of that entity to the entire document text.”

The word “kitchen” is treated exactly the same in both versions of the above sentence, so regardless if “men” or “women” is used, “kitchen” receives a 0 score, 0 magnitude and 0.20 salience grading. That essentially means that the algorithm doesn’t identify any sentiment with the word “kitchen” and that the word itself carries only about 20% ‘importance’ within the context of the entire document or sentence.

How do you think the words “women” and “men” scored in the same sentence where the outcome is the same overall sentiment score? The word “women” receives a score of 0, a magnitude of 0 and a salience grading of 0.80. With regards to sentiment, the algorithm treats the word “women” in the same way it treats the word “kitchen”, saying that there is no overall emotional leaning of the text and therefore there is no overall strength of emotion (neither positive nor negative) within that word. In terms of importance or salience, the word “women” receives a grading of 0.80 - it is strongly required in that sentence in order for the whole sentence to make sense.



The word "men" however, receives a total different grading. Within the exact same context, the word "men" receives a score of 0.3, a magnitude of 0.7 and a salience of 0.80. The word "men" within the same context has a higher emotional leaning score as well as a higher overall strength of emotion score.



The author then rewrote the sentences featuring the word "women" but kept the meaning - the context - essentially the same. The sentiment scores were radically different, with the most negative sentence (!), "must not leave", receiving the least overall sentiment and magnitude scores.

Sentence/ Document	Score	Magnitude
I think women should stay in the kitchen because that's what they're best at *[original document]*	0.8	0.8
Women belong in the kitchen	0.1	0.1
Women should stay in the kitchen	0.2	0.2
Women must not leave the kitchen	0	0

What appears to be taking place is that machine learning algorithms lack the ability to understand context and so certain words are treated differently in what is otherwise the exactly same sentence. The irony here is profound and worth mentioning, as in the API document, the word "context" is given to describe the words surrounding the entities.

In addition, while a person reading the above statements can see how the context, or meaning contained within each statement, is quite similar - a woman's place is in the kitchen - the sentiment scores range from a neutral 0 to a positive 0.8. While machines can be trained to mathematically grade sentiment, it's worth being reminded that they actually don't feel a thing.

Coding with our eyes wide open

Emerging technology like chatbots and voice-activated AI are fast becoming another standard way people can engage with brands and businesses, but as the space between tech teams and the brand teams begins to narrow, there's another that is starting to widen.

Research published in MIT Sloan Review[27] last year demonstrates this. Almost 85% of business executives believe that AI will allow their companies to obtain or sustain a competitive advantage but only about *one in five companies* has incorporated AI in *some* offerings or processes. Read another way, 85% of business executives acknowledge change is coming, yet only 20% have *sort of* done something about it. There is a tremendous gap between ambition and execution.

One potential reason for this gap has to do with complexity and the lack of 'translators' within a business. Understanding AI and machine learning is tremendously complex, with many AI researchers likening it to alchemy. Ali Rahimi, a researcher in artificial intelligence (AI) at Google spoke at an AI conference last year and said that many researchers don't know why some algorithms work and others don't, nor do rigorous criteria for choosing one AI architecture over another exist. "There's an anguish in the field," Rahimi said. "Many of us feel like we're operating on an alien technology."[28] With even the experts grappling with what architectures to use and why some code seems to work and others not, for non-experts, AI and ML can feel like voodoo let alone alchemy. Having individuals within a business who can work with technical teams to ask questions that will generate the kind of answers that can be translated into strategic recommendations is essential to moving forward. However our bias to connect AI to science and therefore scientists may be preventing the business community from employing the kind of diverse thinkers they really need.

While we are making gains on opening up computer science and AI to multiple ways of knowing and understanding through the rise of complementary fields (like Service Design and Human-Computer Interaction), greater adoption of diversity and inclusion initiatives and more attention being given to the marginal user, we still have a long way to go in employing social scientists into data science teams.

To tackle gender and racial inequality in the workplace, a number of initiatives have launched in an attempt to restructure and redistribute power. Programmes that build confidence, generate support and understanding as well as celebrate the power of difference have grown in scale and strength. Within these diversity and inclusion conversations, the subject of neurodiversity is slowly starting to amass support. Neurodiversity celebrates a spectrum of neurological conditions (dyslexia, autism, ADHD, dyspraxia) and acknowledges and includes different ways in which people learn and process information. While in no way should attention be diverted from the meaning and attention being given to neurodiversity, it seems clear that many are starting to openly celebrate different kinds of problem-solving. This could be the catalyst for truly changing the makeup of your data science and AI teams.

A more diverse data and AI team will not only be able to paint a more colourful and interesting picture for business executives in order to make more insightful decisions, but will also have the ability to interrogate data from a number of perspectives. This multi-perspective approach is required in order to spot harmful and damaging bias in data that can damage a business's reputation, dismantle trust with customers and/or reinforce offline bias that supports unfair social practices.

The above-mentioned research from MIT Sloan Review also revealed that those who are leading the charge among the business community have a much deeper appreciation for what is required to produce AI. In addition, businesses who are succeeding are also more likely to have senior leadership who have developed a business case for AI initiatives. This is another *extremely* important point, one that isn't being spoken about enough.

If technology is going to be used to give your business a voice, what do you want it to say?

---

[27] MIT Sloan Management Review (2017), *Reshaping Business With Artificial Intelligence* ,
https://sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/
[28] Science Magazine (2018), *AI Researchers Allege That Machine Learning Is Alchemy* ,
http://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy

Developing a business case for AI initiatives needs to be strongly connected to overall organisational purpose, as your technology - much like an employee - will soon be (if not already are) speaking on your behalf. What is the purpose of your organisation? If you don't know the purpose of your organisation, it's going to be incredibly difficult to identify a use-case for this kind of technology, understand what data is needed and ensure that it executes within the context of the organisation's overall purpose. With so many organisations around the world selling similar products and services, customers are using purpose and values as a way to differentiate organisations from one another. The 2017 Edelman Earned Brand study[29] shows that 50% of consumers worldwide say that they make belief-driven purchase decisions. Belief-driven buyers mean that consumers will buy your brand, switch from it, avoid it and boycott it over your stance on an issue. Further, Edelman Trust Barometer data[30] shows the importance of certain factors in building trust with customers. This includes treating employees well, listening to customer needs and feedback, having ethical business practices, having transparent/open business practices, placing customers ahead of profits and addressing the needs of society in its everyday business.

While not comprehensive enough on its own, coupled with *purpose, intent* is important to consider. What impact on the world do you intend to have? The value of intent is divided into three strands: 1) the actualising of *Purpose,* 2) establishing a baseline in order to measure impact and 3) serving as de facto documentation of business practice in possible, future legal proceedings. It is worth reiterating that documenting purpose and intent is not enough. Firstly, you have to use these two concepts, like lenses, through which you view datasets as you begin the mindful construction of chatbots and voice-activated AI. Secondly, you have to use these two concepts as a baseline to measure what kind of experience people had and what the impact was.

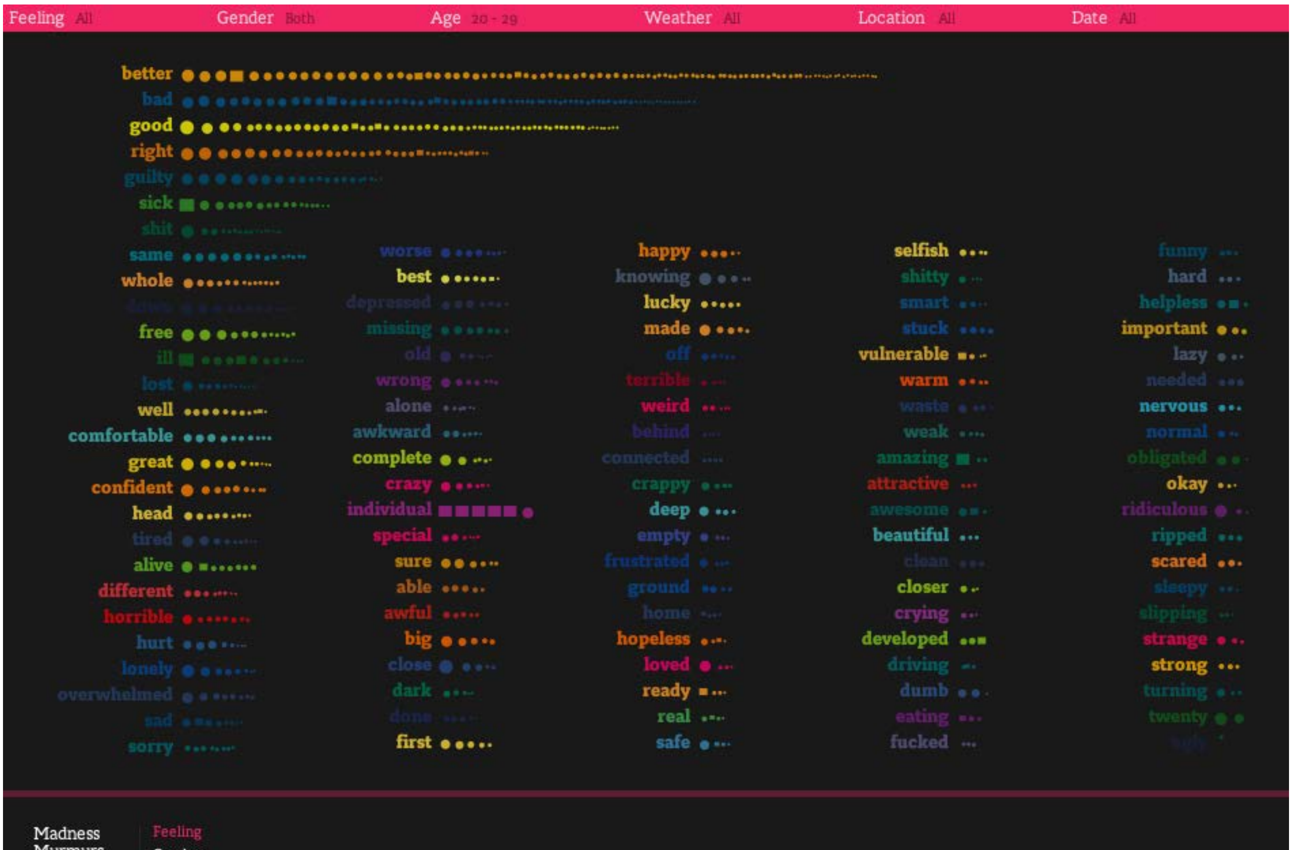
PIIE: A framework for building chatbots and voice-activated AI

Questions surrounding an organisation's purpose and intent are of an ethical nature as they lead to discussions of how one wants to be, act and contribute in the world. These are not easy discussions to have not only because they are dependent on a whole host of factors but also because it is very difficult to get multiple stakeholders agreeing on the same points. Today, nearly every organisation requires technology to function – sold as a product, used to run a service, implemented to connect with individuals internally and externally – it allows for the heart of an organisation to keep beating *and* creates the halo effect that extends and integrates with others. In much the same way the offline has become embedded within the online and vice versa, **to talk about business ethics is now to talk about technology ethics.**

Mandating the use of one ethical theory over another is not constructive as debates over who is right and who is wrong are likely to lead to lots of arguments and little to no action, an ethical dilemma in itself. Like the case of the AI researchers, mentioned above, who are experiencing anguish within their field, academic philosophers seem to be facing similar confusion. The 2009 PhilPapers survey of philosophy faculty questioning which theory of ethics should be developed for machine ethics revealed disagreement amongst the 'experts'. 26% accepted or leaned towards deontology, 24% accepted or leaned towards consequentialism and 18% accepted or leaned towards virtue ethics. The remaining 32% favoured other approaches.[31] As you can see, even the experts don't know. Rather than twisting ourselves into theoretical ethical knots, conversations around purpose and intent can help to incubate ethical action in a way that feels constructive and real to an organisation. After all, this needs to be about action and not just theory. What does your business do and how does it behave towards others? What kind of impact do you intend to have on the world? The answers to these questions

[29] Edelman (2017), *Beyond No Brand's Land,* https://www.edelman.com/earned-brand
[30] Edelman (2018), 2018 *Trust Barometer Global Report,* http://cms.edelman.com/sites/default/files/2018-02/2018_Edelman_Trust_Barometer_Global_Report_FEB.pdf
[31] Bogosian, K. (2017). Implementation of Moral Uncertainty in Intelligent Machines. Minds & Machines, (27) 591 - 608

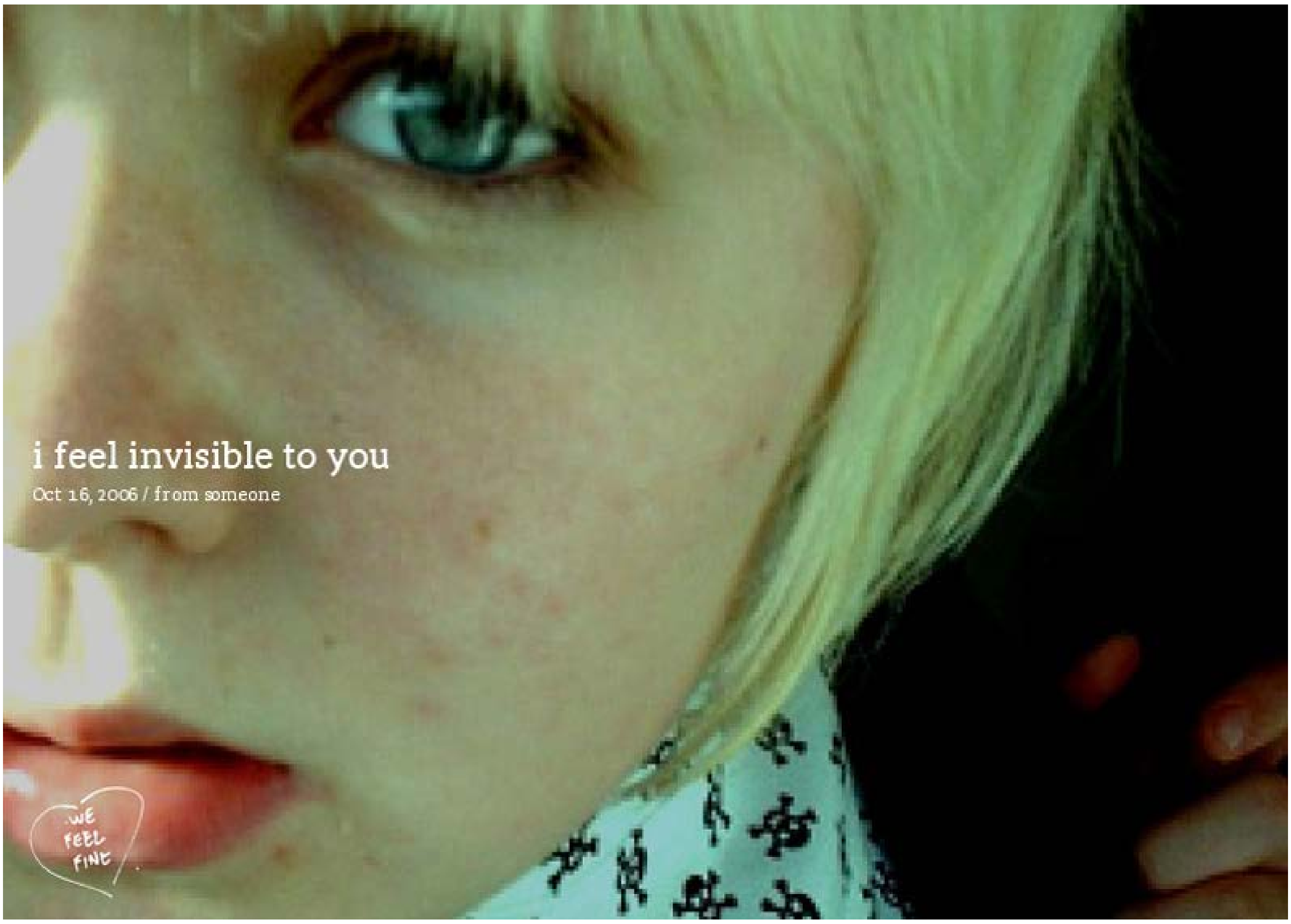
will shape the kind of technology you build by providing a lens through which to make decisions which are right for *your* organisation. Like data, ethics too is personal, subjective and intricately weaved to context. A one size fits all solution, cannot and will not work in practice.
We Feel Fine,[32] an emotional search engine and web-based artwork by Sepandar D. Kamvar and Jonathan Harris, explored the road less traveled with their sentiment analysis tool that aimed to understand more about emotions themselves rather than simply highlighting the connections between topics and emotions. This novel approach to analysing emotions transformed big data into accessible understanding. This quali-quantitative approach centered around emotion helped to inform design considerations of the *We Feel Fine* website.[33] The result was an immersive web experience that helped to shape understanding of emotions through creative displays of scale and cultivate empathy by sharing the emotion expressed within its original context.



Feeling breakdown of feelings from people aged 20 to 29

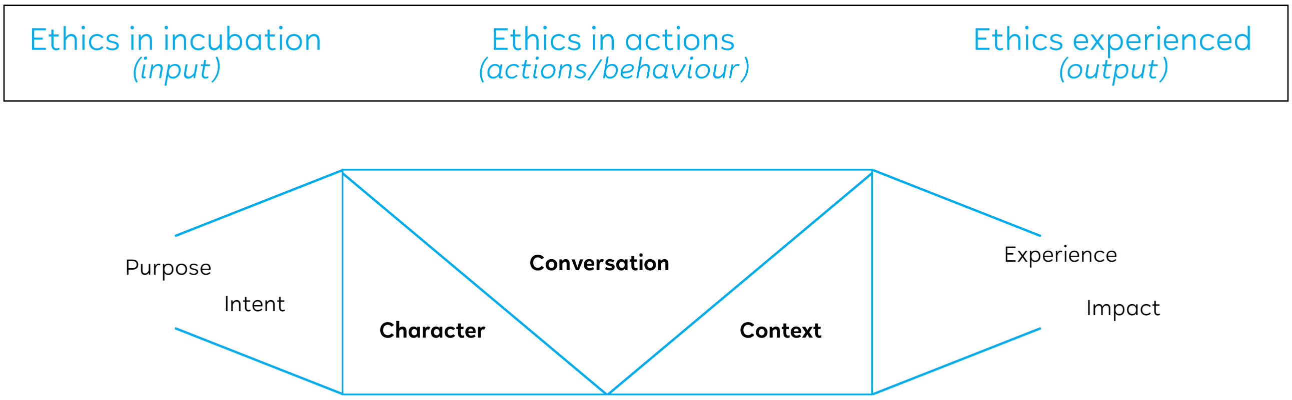
---

[32] Kamvar, S.D & Harris, J. (2011). *We Feel Fine* and Searching the Emotional Web. WSDM'11, February 9 - 12, 2011, Hong Kong, China.
[33] We Feel Fine, http://wefeelfine.org/



Inspired by the quali-quantitative approach expressed in the *We Feel Fine project,* the framework presented in this paper aims to understand more about data itself rather than simply the connection, or vector representation, between words.

Organised around the principles of *Purpose* and *Intent,* the framework allows for, and opens up mindful questioning to (a) the histories and ideologies that underscore big data sets and (b) helps to guide action when diagnosing datasets for suitability and relevance. Rather than critiquing the relationship between different words or scouring datasets with billions of words for (subjective) bias, the proposed PIIE framework transforms a quantitative process into a collaborative, qualitative exercise. PIIE helps to unite the seemingly disparate stages of the tech creation process, within this context chatbots and voice-activated AI, by supporting and facilitating conversations between technologists, social researchers and operations teams through the common language of *Purpose* and *Intent, Impact* and *Experience.* This common language supports individuals as they make the mental leaps required between the intellectual talents of data science, social science and business science.



Executing PIIE during the creation of a chatbot or voice-activated AI

One of the major challenges experienced while working on this project was developing a shared language through which to analyse data. Engineers and social researchers use different language to describe their thinking, methods and analysis processes. This impacts the questions they ask and what they end up 'seeing' in the data.

As a result, we developed a common language through shared definitions and understanding of purpose and intent. Like a translation engine, we found we could keep speaking the language of our disciplines while continuing to 'understand' the language of another discipline.

As discussed above, questions surrounding an organisation's purpose and intent are of an ethical nature. Discussions of who your organisation wants to be, how it wants to act in and contribute to the world is when your ethical principles are in incubation phase. These questions are not just to spark debate, they're also to help an organisation refine and redefine their competitive edge in an increasingly chatty marketplace. If consumers are increasingly buying on belief, being crystal clear on your organisation's purpose and intent becomes your single greatest unique selling point that you're able to channel into the development of character, conversation and context. Quite literally, what are you saying?

It's no longer a leap, but a small side step to see how code becomes an extension of your brand and by default, how this provides the experience people have with your organisation. This experience carries impact. If chatbots and voice-activated AI are set to become the voice of your brand, hacking the code to ensure it aligns with your company's overall brand and set of principles is a way to influence ethics in automation and distinguish one chatbot from the next.

A topline measure of experience and impact, would be to ask customers (through surveys, interviews, focus groups), about their experience and the associated impact to evaluate how clearly purpose and intent were expressed through action.

How the PIIE framework influenced *Charisma.ai*

by Guy Gadney, founder of Charisma.ai and To Play For

To Play For's goal is to create a new form of digital storytelling that is based around characters in stories. Our software platform, Charisma.ai, was designed for this, and as a result it produced a series of very human challenges that, we believe, indicate the direction in which all consumer technology is heading – or should head - in the near future.

Charisma.ai is a storytelling platform powered by artificial technology. Its goal is to create interactive stories where consumers talk directly to the characters, and the characters talk back. The end experience is one where audiences build relationships with the characters and can immerse themselves in the story as if it were real. Imagine a retelling of War of the Worlds where you are Skyping with characters as the Martians land, where your influence can change the course of the invasion, where the story plays out on the media through which we get most of our news. This is the strength of Charisma.ai.

However building Charisma.ai rapidly became more than just a technical challenge. Because we were building a technology that was so human focused – building it both for writers to create stories, and consumers to experience them, and because the platform is so focused on personality, we started to think about its own characteristics and values.

We studied the role of characters in stories: protagonists, antagonists, sidekicks, heroes and villains, and started to imbue Charisma™with these character types. And as we progressed, we realized that we were increasingly weaving ethical and human philosophies into our technological development. We started thinking about gender and diversity in stories. We started thinking about how historical literature and scripts had so much cultural bias in them; that we could not wholly rely on them as a data set to guide our future. And then we started thinking about bias in data sets, and realized that the problem we were encountering was, and remains, a dark shadow in the growth of artificial intelligence.

Our problem was that if we trained our Charisma.ai system on unethical data – say the whole of Twitter conversations – it would become a biased system that did not reflect the reality of humankind today.

Charisma's AI is threaded throughout the platform. It analyses sentences, extracting emotional and linguistic intent so that appropriate responses can be given. It provides context for the characters to exist in – very different from say Alexa or Siri, and enables a conversation to be had with our characters, rather than just a question and answer session. To create this character-strong platform, we became very aware of our own voice as a company and as a team. We questioned and agreed our values – not for some corporate exercise, but because we knew that Charisma.ai would have character, and we wanted it to be a good one.

So as we worked on Charisma.ai, we thought about its principles and ethics. Was our choice of data sets going to create inappropriate content? Would our conversations become to gendered, too cultural siloed? Would we be negatively influencing the creativity of our writers by infusing Charisma.ai with the wrong data?

Our goal is to bring human storytelling to a world of digital, not the other way around. By involving ourselves from the beginning with the ReFig research, we have been able to check our technical assumptions and develop narrative into machine learning without losing the human connection.

Our view is that stories will always need human creativity. A goal to automate story creation using AI shows a profound ignorance of both technology and its purpose in society. **Our purpose is provide a system which strengthens writers and allows them to create and pioneer new forms of storytelling for new audiences.** By help shape this research - primarily for the video games industry through ReFig – we hope to have made it relevant for a broad range of consumer digital media organisations who are wrestling with their own identity in the age of the digital persona.

A systems approach to tackling bias and considering ethical issues

As well as looking at internal practices to identify and overcome bias within AI, organisations need to think about how they start to build this capability externally. Broadly speaking, the tech industry's capability to talk about bias, identify it in the datasets it uses and then make decisions is still developing. And as the predominant producers of knowledge around AI, this lack of capability has long reaching and complex consequences. At a recent Nesta event on AI and the future of work,[34] panelists discussed ethical, social, cultural and philosophical issues – such as the long term impact of work being distributed by algorithms (i.e. an app decides which taxi answers the customer's need) – concluding that many products and tech companies are unable to grapple with these issues. This view reflects the fact that over the past 12 months ethical issues have been raised repeatedly at conferences and in media, without any reliable solutions or industry standards being agreed to deal with them.

Recently, organisations like AI Now in the US and Nesta in the UK have released the beginnings of best practice frameworks around machine learning and algorithmic decision-making.[35] Furthermore, the International Standards Body is starting work to establish an ISO on a framework for AI systems using machine learning.[36] While these are encouraging, they are only very recent examples and we are yet to see their impact. In addition, these examples are largely led from

[34] http://www.nesta.org.uk/event/futurefest-forward-ai-and-future-work
[35] Algorithmic Impact Assessments, AI Now, April 2018 and Code of Standards for Public Sector Algorithmic Decision Making, Nesta, February 2018.
[36] ISO/IEC AWI 23053
Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML), https://www.iso.org/standard/74438.html

an academic-institutional perspective, rather than based on a shared consensus of what works across the tech industry. The sociotechnical blindness we have also leads us to thinking that we need to come up with new ways of dealing with bias. However, as Joanna J. Bryson has said repeatedly, we already have established ways of dealing with bias in society and it's time we bring these to bear on the tech industry rather than treat them as needing a new set of rules.[37]

An ecosystem of industry, academic, government and community members is the best way to bring together the different knowledge needed to guide the ethical development of AI. We need to engage these networks with the task of identifying techniques and case studies that can help to lift our collective understanding and ability to address these issues. Open source tools and techniques, industry platforms and communities of practice, need to flourish around the questions raised in this report. Universities have started embedding ethics in their computer science courses[38] and we need to have an ethical ecosystem in place for these graduates to continue developing their skills and knowledge in this space once they join the industry and begin building things.

R&R: Recapping Recommendations

1. Critically discuss your organisations purpose and intent
- ➤ What is your organisation's purpose and how does it achieve this purpose?
- ➤ What kind of impact does your organisation intend to have on the world?
- ➤ How does your organisation behave towards others? How do you measure the impact of that behaviour?

2. Diversify your team
- ➤ Do we have a balance between people who are interesting in knowing and understanding the world of machines and the world of humans?
- ➤ Do we really need another data scientist? Do we need a social scientist instead?
- ➤ Do we have a translator in our business that can act as a mediator between operations and strategy?

3. Interrogate the data
- ➤ Where did the data come from?
- ➤ What or who is included in the data, and what or who is excluded from the data?
- ➤ Is the tone and language consistent with your organisation's purpose and brand guidelines?

4. Consider the marginal user
- ➤ Who are you designing your service or product for? Can you identify a marginal user who is currently underserved in this area?
- ➤ What are their specific needs and what are the specific barriers and pain points that face them? What are their specific strengths and viewpoints?
- ➤ What are the different participatory methods you practice so that your marginal user can co-create or have direct input into the development of your product or service?

5. Establish industry partnerships
- ➤ How are you learning from the industry to identify and address bias present in your AI-powered service or product?
- ➤ How are you contributing your own knowledge about this into the industry?
- ➤ What aspects of the work of your organisation can be conducted using 'open innovation' methods, so you can draw on and learn from a range of views across society?

[37] Bryson, J. J. (2018) "Patiency is not a virtue: the design of intelligent systems and systems of ethics", Ethics and Information Technology, Vol 20, Issue 1, March, pp. 15-26 (https://link.springer.com/article/10.1007/s10676-018-9448-6)
[38] Singer, N. (2018) "Tech's Ethical 'Dark Side': Harvard, Stanford and Others Want to Address It", New York Times, 12 February. Available at: https://www.nytimes.com/2018/02/12/business/computer-science-ethics-courses.html, Accessed 18 May 2018.

Conclusion

Innovation within the emerging technology industry is an expensive endeavour and as a result, we're seeing more developments taking place within the private sector rather than the public and academic sectors. However, non-profit institutions are leading the charge with regards to critical thinking and raising awareness of technology's impact on society. The people in the middle, the masses, are caught in the middle oscillating between confusion and fear, desire and need. Between it all, purpose seems to be standing out.

The rise of purposeful business, millennials and generation Z employees searching for purposeful careers and customers, empowered through choice, are now increasingly searching for products and services that match their belief. The language of ethics while inspiring and meaningful, can prove as complex to navigate as lines of machine learning code. But purpose, that's a word we all understand.

The Research Team

The research was led by digital sociologist Lisa Talia Moretti with contributions from research assistant feminist AI researcher Josie Young, and founder of Charisma.ai Guy Gadney.

The team would like to extend their thanks to senior software developer Stephen Reid for his valuable efforts in furthering this research.

The team would like to extend their thanks and gratitude to ReFiG for making this research possible. Refiguring Innovation in Games (ReFiG) is a 5 year project supported by the Social Sciences and Humanities Research Council of Canada. Composed of an international collective of scholars, community organizers and industry representatives, ReFiG is committed to promoting diversity and equity in the game industry and culture and effecting real change in a space that has been exclusionary to so many.

Thank you to Dr Sarah Atkinson at King's College London for the opportunity to work on this project.

Lisa Talia Moretti - https://www.linkedin.com/in/lisataliamoretti/
Josie Young - https://www.linkedin.com/in/josieswords-young/
Guy Gadney - https://www.linkedin.com/in/guygadney/
Stephen Reid - https://www.linkedin.com/in/steveyblam/

Art Direction and design: Lucia Viglietti - https://www.linkedin.com/in/lucia-viglietti/

http://www.refig.ca/