CaladoF

Table of Contents

[1. thank you](#orgc1ce09e)

[2. 1. overview of diss - critical look at technology - 10](#orgeb62c1e)

[2.1. main idea](#org1bda688)

[2.2. orlando close reading](#org9d549ae)

[2.3. reduction](#org6ad1c2f)

[3. detail of my work - text encoding - 20 min.](#org7b7f3ca)

[3.1. queer studies trajectory](#orgb8b03a4)

[3.2. text encoding chapter overview](#org58f37c7)

[3.3. hierarchy](#org7ba4df8)

[3.4. Dominance structures in the archive](#orgd89a171)

[3.5. qtt](#org78115c0) and hands-on experimentation

[4. 3. where my work is now going - algorithmic bias - 10](#orge704b12)

[4.1. anti trans & text generation](#org6998c4f)

# 

# thank you

Thank you for having me. Today I'll be talking about my dissertation, which I defended in November in the English department at the Graduate Center at CUNY. For that project, I applied a methodological approach based in Queer Studies to analyze Digital Humanities tools that work with language and language forms. I will show how these digital tools flatten or reduce some expressive qualities of language, particularly language that relates to sex, gender, and sexuality, and how Queer Studies frameworks can be deployed to engage with these qualities. Toward the end of the talk, I will discuss how this work has led me to a follow-up project, which applies the same analytical spirit to examine AI tools and gender bias, specifically by using AI to analyze legislation that limits transgender peoples’ rights, legislation that is currently proliferating across the United States.

# 1. overview of diss - critical look at technology - 10

## main idea

First, my dissertation, entitled “Since No Expressions Do: Queer Tools for Reading Literature”, uses literature, and queer literature in particular, as a test case for exploring what happens to language when it is transformed into electronic formats and processed with digital tools. Each of my chapters takes up a different aspect of computation–such as programming logics, data structures, and interface effects–to explore how that aspect handles verbal forms of sex, gender, and sexuality. I find that digital processes, which by design flatten the semantic expressiveness of language into text data that can be categorized and counted by the machine—that these same processes can be reframed to surface queer gender and desire in text.

My methodology is to do close readings of digital tools, to examine how they handle the complexity and nuances of language forms. I look into the mechanism of each tool, tracing the way that it transforms language into text data for electronic processing and display. I read the logical structures of computer code, across programming and markup languages such as Python, JavaScript, and XML. From computer code’s internal logics, I isolate a specific constraint, an aspect that collapses the expressive qualities of language into computable data. Then, I draw from theorizing in Queer Studies and its critiques to find ways of working within this constraint. Queer Studies offers this project methodologies for recognizing and resisting totalizing and oppressive systems, such as that of heteronormativity. I look to these critical frameworks as models for reading the strict logics of computer code.

## orlando close reading

Before elaborating on my theoretical intervention, I'll share a small example of what this close reading method looks like, to concretize the kinds of reductions that are inherent to computational processes, and how these reductions affect language data. Drawing from Literary Studies, close reading is a method of analysis that pays attention to textual detail. It focuses on specific words, phrases, and punctuation for interpreting the tone, mood, and meaning of passages. It is a way of grounding a reading of a text in language forms.

SLIDE 2: Tilda Swinton image from *Orlando* film.

I’ll begin with a close reading from the first sentence of Orlando, a novel by Virginia Woolf. And here’s an image of a famous film adaptation, by Sally Potter, of the novel. For those who are unfamiliar with this text, it tells the story of Orlando, a 16th century English nobleman who has a sex change halfway through the book, and lives the rest of her life as a woman.

SLIDE 3: first sentence of Orlando.

The story opens with a scene of racialized violence, a scene that embeds important assumptions about the construction gender:

“He–for there could be no doubt of his sex, though the fashion of the time did something to disguise it—was in the act of slicing at the head of a Moor which swung from the rafters” (11).

In this opening scene, a teenage Orlando swipes at a severed human head that his ancestors brought back to England from the Crusades. Within this backdrop of imperial conquest, this first sentence makes some qualifications about gender, qualifications that are precariously held up by subtle details in the text: first, the reassurance that there could be “no doubt of [Orlando's] sex,” belies itself: it doesn't serve to reassure, but rather, is conspicuous in its certitude. That this reassurance follows an em dash interrupting the first word, a bold gender declaration–HE–draws even more attention to gender as something that is in contention. The associations here deepen when we connect this certainty, and even defensiveness, about gender with the imagery of racialized violence, the “slicing at the head of a moor.” Masculinity here appears to be co-constructed with nationality and race, which heralds a central tension in the ensuing story.

If one were to analyze this text with text analysis tools, using the Python programming language, for example, the transformation of the text into a dataset would evacuate the ways that gender is unsettled in the sentence. That is because all text analysis is based on counting: counting the number of times a certain word appears, and how many times its surrounding words appear, and what words tend to surround those words, and so on, with the ultimate goal of deriving patterns from these word frequencies. In order to count the words, however, the text must be prepared: Semantic features which are deemed minor and frequently occurring (like articles, prepositions, punctuation) must be removed. This process of “text cleaning” ensures that minor details won't skew the counting, the word frequencies, of the text, which is the basic foundation of all text analysis tasks. For this sentence in particular, that means the pronoun “HE” and the em dash following that “HE” would be removed, so that the sentence would appear as this list of words:

SLIDE 4: list of tokens of the first sentence after cleaning

‘could’, ‘doubt’, ‘sex’, ‘though’, ‘fashion’, ‘time’, ‘something’, ‘disguise’, ‘act’, ‘slicing’, ‘head’, ‘moor’, ‘swung’, ‘rafters’.

While this process makes the text amenable to quantitative analysis, creating a computable dataset of words, it also loses how pronouns and punctuation affect the construction of gender in this sentence.

## reduction

Quantitative methods for analyzing language, also known as “Distant Reading” methods, can lead to interesting and provocative results. By stripping minor details, connections between more substantial words and language patterns might emerge. My work is not intervening in those kinds of analyses. Rather, I’m interested in reading the reductions required by computational processes. I see these reductions starting points for thinking through language as an enabling constraint in human computer interaction.

I turn to Queer Studies, which arose in the early 1990s, in the wake of the AIDs crisis, and offers methods for understanding (and revealing the contradictions of) sex, gender, and sexuality as social constructions. Because Queer Studies seeks strategies for not only recognizing oppressive structures, like compulsory heterosexuality, but of working within and against them simultaneously, it offers useful models for thinking about the ways that machine processes create a compulsory disambiguation of language data.

For example, with text analysis methods, I make a connection between the processes of cleaning a text with the way that gender is defined according to Judith Butler's famous theory of Gender Performativity, a theory that helped to inaugurate the field of Queer Studies. This connection centers on the concept of iteration, which is central to both text cleaning and gender performativity.

For text cleaning, I examine how the Python programming language works with text data through a mechanism known as the “loop,” which is a core building block of algorithms. The loop, technically called the “for-loop” enables a program to do things en masse to data, by repeatedly applying the same series of actions to each piece of data, one by one. The technical language describes loops as a process of “iterating” over data.

SLIDE 5: definition of iteration from python.org

Here is a definition of the loop from the Python wiki.

“... loops are used when you have a block of code which you want to repeat a fixed number of times. The “for-loop” … iterates over the members of a sequence in order, executing the block each time” (https://wiki.python.org)

I then turn to Judith Butler’s concept of gender performativity, which defines gender as a series of repeated acts, of behaviors, that conform to an authorizing norm. Her central intervention is that behavior determines identity, or at least the illusion of identity. According to Butler, it is by adopting gendered modes of behavior that subjects come to exist at all.

SLIDE 6: definition of performativity by Butler.

Here is an excerpt of Butler’s definition of Gender Performativity:

“Performativity must be understood not as a singular or deliberate “act,” but, rather, as the reiterative and citational practice by which discourse produces the effects that it names” (Butler 2)

From this definition of Gender Performativity, I am interested in the characterization of gender as “re-iterative” practice. As Butler explains later, this practice is the key to resisting gender norms from within: by repeatedly performing gender constraints in ways that deviate from gender norms, for example in drag performances, subjects can subvert social expectations of gender.

Taking this shared quality of iteration between Python and gender, I propose a text analysis practice that combines different views for analyzing text. I demonstrate this process in a custom text analysis application, which I designed as part of a software package, “The Queer Text Toolkit,” that is supplementary to my written dissertation.

SLIDE 7: screengrap of queer text toolkit

Here you can see some screen grabs of the text analysis application, which runs on the command line, and uses Python libraries for text analysis, like NLTK and Word2Vec, for those who are familiar with Natural Language Processing. In contrast to traditional so-called “reproducible” social science methods for text analysis, my custom application, which is available for download on github, offers an iterative method that flips between close, textual detail and distant views of language data.

# detail of my work - text encoding - 20 min.

## queer studies trajectory

Now that I’ve shown a small example of my analytical approach, I’ll discuss Queer Studies as a field and my positioning within it.

Throughout my dissertation, my thinking about Queer Studies reflects the trajectory of its arguments over time: while the field began by disentangling the contradictions within gender, sex, and sexuality within hetero-patriarchy, it changed with the queer subject’s increasing incorporation into neoliberal politics. With the rise of progressive agendas, such as the prioritization of same-sex marriage and inclusion in the military, Queer Studies had to reformulate its relationship to transgression. Numerous critiques emerged to resist trends toward normativity (sometimes called “homonormativity”). On the one hand, there was an embrace of negative affective modes such as shame, failure, despair, and loss as constitutive of Queer subjectivity. By embracing negativity, without trying to recuperate or fix it, one keeps Queerness as something that cannot be captured or redeemed. On the other hand, there was a doubling down on transgression at all costs, which manifested in theorizing around a total rejection of mainstream acceptance, such as in the family and in so-called “reproductive futurism,” a rejection that is known as the “the anti-relational” or “anti-social” turn.

In response to these developments, one critique, called “Queer of Color Critique,” opens the purview of Queer Studies in a way that influences my approach toward analyzing technology throughout my project.

SLIDE 8: Cruising Utopia cover

I turn to Jose Esteban Munoz, who brings intersectional thinking from Women of Color Feminism in the 80s and 90s to re-energize Queer's political potential. As a corrective to Queer Studies' implicit (and sometimes not so implicit) whiteness in its critical frameworks, where gains in the Queer community are trapped within what Munoz describes as “the limiting, normative time of the present,” he offers a vision for queerness as futurity.

SLIDE 9: Munoz’s quote

He explains that,

“Queerness is a structuring and educated mode of desiring that allows us to see and feel beyond the quagmire of the present… Queerness is a longing that propels us onward, beyond romances of the negative and toiling in the present. Queerness is that thing that lets us feel that this world is not enough, that indeed something is missing.” (Cruising Utopia)

Munoz’s thinking opens possibilities of theorizing from political and social structures that allow some subjects visibility and acceptance while attempting to erase the existence of others. Queerness conceptualized in this way, as a desire for what is not present, pushes Queer Studies beyond the boundaries of sex, sexuality, and gender and into its imbrications with other oppressive structures, like race. In my dissertation, this move in Queer of Color Critique encourages my continual re-thinking of how computational structures constrain language into computable elements, into components that function within larger, totalizing systems.

## text encoding chapter overview

I’ll provide an example from another chapter of my dissertation, where I draw from critical methods in the history of the Black Atlantic to engage with the constraints of data formats.

SLIDE 10: XML

In this chapter, I consider a data structure called XML (short for eXtensible Markup Language), which is widely used for encoding, or “marking up” written content within a hierarchical structure, very similar to HTML.

SLIDE 11: XML hierarchy

Like HTML, this data structure uses “tags” to wrap written content. Here, the text is written in gray, and the tags appear in blue. The tags have descriptive labels like <page> for a page, <line> for a line of text, and <quote> for a quotation.

According to XML requirements, all written content must be enclosed with these descriptive tags, which must themselves be nested neatly within one another, so that so-called “child” elements are totally contained within their parent element. Due to this nested format, XML code resembles a tree structure, with one “root” element, and nested within it, a number of branches.

Like with text analysis, I was interested in locating XML’s constraint—how it reduces expressive aspects of gender, sex, and sexuality in language. To seek out this constraint, I used it in an editorial project, for transcribing the written text of a literary manuscript that was edited multiple times to remove evidence of homosexuality.

SLIDE 12: Dorian Gray MS & Wilde

I chose the manuscript of Oscar Wilde's novel, The Picture of Dorian Gray, which Wilde revised heavily before its first publication in 1890. For those who don't know the story, it takes place in late 19th century London, where a beautiful youth named Dorian Gray makes a deal so he can stay forever young. While Dorian keeps his youth, his portrait, painted by his artist friend named Basil Hallward, would grow old and ugly in his place. The story then follows Dorian's descent into moral abandon, through which he eludes the marks of age, remaining apparently untouched by corruption. At the time of its publication, the novel was variously criticized for its ambiguous stance toward Dorian's hedonistic lifestyle, as well as undertones of homoeroticism, which emerge in its various exhortations of masculine youth and beauty.

The critical scholarship on the manuscript explores how Wilde’s revisions stifle suggestions of homoeroticism. Wilde’s strategy for doing so, critics generally agree, is to transform erotic passion into a focus on art and an appreciation of aesthetics. The changes that Wilde makes to the text are often small ones, but they have significant effects on tone, mood, and connotation.

SLIDE 13: image from MS with diplo rendering of page 9

In one example, Wilde makes subtle changes to mute the intensity in the dialogue, replacing tension with a sense of levity. In this scene, the painter, Basil Hallward, is about to explain to his friend, Lord Henry, why he cannot exhibit his portrait of Dorian Gray. I will read the text, prior to revision:

Lord Henry hesitated for a moment. “And what is that?” he asked in a low voice.

“I will tell you,” said Hallward, and a look of pain came over his face.

“Don't if you would rather not,” murmured his companion, looking at him. (9)

In the revision, Lord Henry’s hesitation becomes a laugh, and Basil’s look of pain becomes an expression of perplexity. These changes trade affliction for confusion, and diminish the fatalistic connotations of Basil's passion. They also suggest an underlying sense of shame or restraint, especially from Lord Henry, who no longer hesitates or speaks in a low voice, but instead encourages Basil to share his story. Together, these changes work to lighten the mood of the scene.

To encode these revisions from the manuscript into the XML data format, I customized a variant of the XML language. Using the XML as a base, I created a schema with custom attributes for registering Wilde's revisions according to thematic categories, like “passion,” “intimacy,” “fatality,” and “beauty.” I chose these labels for the way they work together to aestheticize Dorian’s character, turning him from an erotic object into an aesthetic one, and to diminish Basil’s consuming passion for him. In addition to theme, I also wanted to markup the text in a physical dimension, to help indicate each instance or round of revision. So I created an attribute to mark the pen strokes obscuring each piece of deleted text.

SLIDE 14: XML of page 9

This image of XML code here shows how I encoded the same passage from the manuscript. We see the tree-like structure of the XML, along with the tags (in dark yellow) that are particular to editorial projects, like tags for indicating deleted text in <del>, added text in <add>, and their parent tag, <mod>, which stands for modification. Additionally, we can see attributes in green, like “type” to indicate the type of modification (in this case, a substitution of one text for another), “rend” to indicate how a bit of deleted text should be rendered, such as with a strikethrough; and “place” to indicate where the addition was made, which is above the line. Finally, we also see my custom elements: @theme and @strokes, to indicate the category for revision and the number of pen strokes that Wilde used on the section.

## hierarchy

My work on the minute, detailed labor of encoding the manuscript reveals a constraining quality about the XML data structure—that of hierarchical dominance. The XML requires that all data be contained as discrete components within its bounded structure, which cannot overlap unless the inner element is fully nested within an outer element. When child elements do not fit neatly within the parent, the file will display error.

Due to this constraint with overlap, in Wilde’s manuscript, it was difficult to indicate the different rounds of revision with accuracy. For example, one page of the manuscript shows what are apparently multiple rounds of revision, that together suppress intimacy and passion in the text.

SLIDE 15: image of MS page 20 with transcription:

In this scene, the painter explains the effect that Dorian Gray has on his art:

“You remember that landscape of mine… It is one of the best things I have ever done. And why is it so? Because, while I was painting it, Dorian Gray sat beside me, and as he leaned across to look at it, his cheek just brushed my cheek. The world becomes young to me when I hold his hand, as when I see him, the centuries yield up all their secrets!”

His friend’s response, as you can see, is mostly illegible, but I could make out some words and phases, like “you must not talk”, “his power,” and “slave”, as well as the final line, which reads: “It is worse than wicked, it is silly. I hate Dorian Gray.”

In revising the passage, Wilde replaces “cheek just brushed my cheek,” with “hair just touched my hand,” before striking out the entire line, as well as the following sentence. He also strikes out much of Lord Henry’s response, replacing Lord Henry’s intensity with a relatively subdued interest when he says, “this is quite wonderful. I must see Dorian Gray.”

SLIDE 16: arrows showing areas of overlap in MS

A span of text, indicated in blue, creates a problem for the encoding. Here, Wilde layers the revision of “cheek just brushed my cheek” with “hair just touched my hand”, within a larger revision that deletes the rest of the paragraph. With the XML, it is very difficult to disentangle these layers while keeping the integrity of each revision as its own entity. For example, here is an encoding that shows the two layers of revision, which keeps the integrity of each layer of revision in tact:

SLIDE 17: encoding of flat revision

Here, the second layer of revision, in the block of text which is most indented in the code, is nested within a larger span of deleted text. However, the hierarchical format of the XML requires more precision, which leads to a problem of overlap. Because Wilde, in this case, struck three individual words, “cheek,” “brushed,” and “cheek” over a longer span of deleted text, each of those words has to be encoded within separate elements. The problem is that this encoding breaks up the longer phrase, the revision as a single unit, into pieces.

SLIDE 18: encoding of deep revision

It splits what was in fact one revision into separate ones, as indicated in this image on the right. Here, there are three modifications, or <mod> elements, within the larger deletion element. This means that any other program used to work with the XML, to search the text or to display it, for example, would treat this revision as three individual changes to the text.

Within my customized schema for marking up this text, the XML’s data model, and its need for strict disambiguation, makes it impossible to mark up the revisions as holistic units. The hierarchical structure, which requires that elements be completely nested within one another, cannot register the revisions as layers, or rounds of revision. The structure requires granularity, which makes it difficult to draw meaningful connections between the rounds of revision, and Wilde’s intentions with each round. As a labelling tool, the XML thus surfaces moments where queerness, in this case, the suppression of queerness, transgresses the bounds of the data structure.

## Dominance structures in the archive

The more that I work with the XML, the more I am interested in the hierarchical data model, and the kinds of dominance structures that are contained within it. There is an underlying lesson in power here, about who has it and who is subject to it.

In Wilde’s text, for example, the dominant power is that of the writer himself, a writer with nearly every privilege—racial, gender, financial, educational, cultural—who censors his own writing. There is also the power that I have, as an encoder that makes decisions and customizes her own schema to mark up this text. What about texts whose writers or subjects are subscribed to larger systems of domination?

SLIDE 19: Wicked Flesh cover

This work with the XML turned my attention to other projects dominated by hierarchical power structures, such as this historical project by Black studies scholar Jessica Marie Johnson. This project is a history of Black women in the 17th and 18th century Atlantic world. Johnson’s book, *Wicked Flesh*, works from archival records like birth certificates, marriage licenses, censuses, and travel accounts. These archival records, which are all written by slave-owning men, traders, or colonial officials, contain minimal information about the lives of Black women, only what these officials deemed necessary to record. From this sparse and highly constrained dataset that, on its own, tells a story of bondage and subjection to power, John constructs vivid scenes that foreground marginalized figures who negotiated their own freedom practices.

SLIDE 20: catti quote

For example, she illustrates a dinner party hosted by Seignora Catti:

“A wealthy merchant in her own right, [who] had leveraged her status as the wife of a European against her commercial savvy and the opportunities and experience of living in the middle ground between the Atlantic Ocean to the west and Wolof sovereignties in the east for her own benefit. Seignora Catti secured property, insider commercial knowledge, and political power.”

As impressive as Seignora Catti sounds, there appears to be no historical writings that feature her name, except as a minor character. A glance at the notes reveals that the source material for Catti’s story comes indirectly, from biographical writings about a commercial agent, named Jean Barbot, who was based in Senegal.

SLIDE 21: notes on barbot

Here is a list of historical and critical writings featuring Barbot, who was apparently an important figure in the French/Senegambian slave trade. In Johnson’s narrative, however, Barbot’s role is delimited to a supporting character, to a guest at Catti’s dinner party whose presence serves to bring Catti into the foreground. In the same vein as other scholars of the Black Atlantic, like Saidiya Hartman, Johnson works within these constraints of indirect and sparse sources to weave a story from the margins of the archive.

Thinking back to the XML, I imagine this story, and others from her book, functioning like the overlap between XML elements. Stories that go beyond the bounds of what is admissible within the data format, by the precision that such a format requires. Spanning numerous fragments, I doubt that the facts of Catti’s story can be traced to its source material in direct lines, in a one-to-one correspondence. Rather, I imagine that they emerge from a kind of gap in between sources, a gap where historical research blurs with creativity.

SLIDE 22: johnson on historians

And at the end of her book, Johnson suggests as much, when she asserts that:

Historians, bound by archives, may scrape dusty folios for sources, may question whether women and girls will appear or worry that when they do appear, they emerge as legends, myths, and motifs representing more than themselves. That is not the intellectual tradition this book was written in.

Rather, Johnson’s historical project requires more than assembling what survives; it requires narrating from what she calls the “deep well,” where “history becomes memory.” It requires not just collecting and describing the fragments, but also the gaps in between them.

## queer form and qtt

My work in the dissertation, as a whole, seeks out such gaps, which enable me to figure the contour, the boundary around the elusive identities, repressed desires, and other coded elements of queerness in text.

I find that the constraints in technological processes and codes are opportunities for reading the workings of power and other oppressive structures, and in that reading, revealing not only what has been flattened or reduced by the machine, but also an incommensurable element, which cannot be measured or captured. Whether these elements are the ways that punctuation and pronouns affect meaning from a “distant reading” vantage point, or whether they operate in overlapping and permeable ways on manuscript pages.

SLIDE 23: image of Queer Text Encoding

I tried to demonstrate these ideas in practice with “the Queer Text Toolkit”, which I mentioned earlier in the context of text analysis. Here you can see a screen grab of the second application from the toolkit, which I created and styled to visualize the XML encoding and the theme of revision from Wilde’s manuscript. The goal with these digital projects, in addition to giving myself a chance to gain technological skills, was to seek a hands-on methodology for putting into practice the theoretical work of the dissertation.

# 3. where my work is now going - algorithmic bias - 10

Now I turn to my follow-up project, which I’ve begun researching as part of my current work at the Princeton University Library.

This project applies a similar method of “close-reading” technical mechanisms toward a new object: Large Language Models made with Artificial Intelligence. My interest is in how these models perpetuate gender bias from their training data onto their outputs. I am currently the middle of a data gathering and cleaning process with the goal of fine-tuning language models to study how they perpetuate gender bias from their training data.

Specifically, I am working with definitions of gender and related terms from legislation that limits trans people’s rights, known as the “anti-trans” legislation, which is currently being debated and passed throughout the US.

SLIDE 24: anti-trans legislation tracker map & stats

Here, on the left, you can see a map of where the bills are most concentrated, and on the right you can see the incredible increase in bills in the last four years. Although the last year, 2023, was a record setting year, the current year, 2024 is rapidly catching up, and we are only in the second month.

## anti trans & text generation

With this project, I am specifically interested in how the anti-trans legislation is being used as a galvanizing tool, as a way of conscripting supporters who may hesitate to condemn homosexuality into a movement that is ultimately phobic of all non-normative identities. As trans activists and scholars have been saying for years, the current focus on trans is only the beginning. Because once this category is sufficiently legislated, the focus will move on to others.

I think it's important that we understand, today, the threat of gender transgression. I'm interested in examining the language outlawing gender transgression, and this particular transgression, of transitioning from one gender to another, or of opting out of binary systems of gender. Why is this particular kind of transgression so controversial, so contentious? why is it so attractiveto much of the public, and used to corral them together?

SLIDE 25: list of gender and related terms

Here’s some examples of the language that I extracted so far from the legislation. I am focused on definitions of gender identity and related terms, like “biological sex” and “gender transition procedure,” which are underlined here.

SLIDE 26: blue text

Here, the text in blue indicates words and phrases that seem to suggest underlying assumptions about binary gender and gender transgression. In particular, I am fascinated with how these assumptions are being constructed in subtle ways, in seemingly harmless formulations. For example, I am interested in the word “regardless,” which appears in many of definitions, and is used to contrast an essentialized notion of binary gender against gender as expression. What does it mean to partition some aspect, like biology, when considering another, like gender presentation or identity? What do such formulations tells us about the allure, the threat, the “seduction,” as Trans Studies scholar Cassius Adair puts it, of gender transgression?

Eventually, I will use these terms to train language models, with the goal of finding out if there’s anything generalizable about these definitions. Leaving aside all the hype about AI, and whether or not they it is really “intelligent,” or moving toward what the industry calls “general intelligence,” AI tools like large language models are really good at one thing: generating *plausible* text. This is a fascinating phenomenon, because it makes them very good at creating content, at improvising, but not at all good at being *creative,* at innovating. It cannot explain, for example, why a particular metaphor is powerful. Trust me, I’ve tried to ask it.

The reason is because language models are “synthetic text generators,” in the words of Emily Bender, one of the authors of the “Stochastic Parrots” paper, a foundational paper in the filed of AI Ethics.

SLIDE 27: Emily bender, stochastic parrots

A language model can only generate what it has already seen before. Even a phenomenon like “hallucination,” that a language model spews text that has no bearing in reality, is based on the tendency of models to reproduce what they've already seen. They hallucinate not because they are creative, but because they are designed from statistical processes to generate what is most plausible.

Ted Chiang, the science fiction writer (author of the story that inspired the film called “Arrival”) offers a useful analogy for understanding how these tools determine plausibility. He describes it like a blurry photograph created by compressing a photo into a smaller format. In a compressed photo, the software guesses what color pixel to use through estimation and calculation. He describes this guess as a kind of “interpolation,” and explains that it works by, “estimating what’s missing by looking at what’s on either side of the gap.” In other words, compression works by looking at the nearby pixels and calculating the average of those pixels to make an inference.

SLIDE 28: parrots!

To demonstrate this effect, I’ve taken a screenshot of the parrot on Emily Bender’s shoulder from the previous slide, and saved it in an uncompressed and in a compressed format. See if you can tell which is which!

The compression process means that the tool is guessing, estimating. What we might call an educated guess. But is guessing the same as creativity? What about improvisation? These are larger questions looming in the background of my research.

Within the blur between inference and creativity, is an interesting place to study the construction of gender, sex, and sexuality. What can we distil, compress, or create about these terms by using large language models? By training a language model to generate the language of discrimination and oppression, this work is part of a larger attempt, as Judith Butler puts it, to “begin, without ending, without mastering, to own—and yet never fully to own—the exclusions by which we proceed” (25).

Thank you.

FEEDBACK:

* ~~Slow down~~
* ~~Explain close reading~~
* ~~Add context for violence at the start~~
* ~~More signposting~~
* ~~Siraj: Emphasize where the problem with the “digital” is~~
  + ~~What is my intervention? What is the effect of XML? Did it work?~~
  + ~~What do I mean layering syntax and meaning?~~
* When answering questions, be practical. Walk them through answers with examples. It’s information studies.
* ~~Visual dimension: add aides for Orlando, Wilde, Bender, Erin Reed.~~ 
  + ~~Use colors for definitions~~
  + ~~Don’t leave Wilde up on the screen when I’ve moved on.~~
* Matt: why is exploring data from a humanities angle important to students of information?
  + What does the literary/sex/gender allow me to bring into Info studies?
* Add more work in Queer DH. Gesture toward it at least, for the Q&A
* What about irony?