Gender Equality: Reimagining our Future Through Art and Technology

5 GENDER EQUALITY

Georgia Tech

The Kendeda Building for Innovative Sustainable Design

Serve-Learn-Sustain Women's Resource Center

About the Exhibition

As part of the 2030 Agenda for Sustainable Development, the United Nations developed 17 broad and interconnected Sustainable Development Goals (SDGs) that address the global challenges humanity faces, such as ending poverty and hunger and reducing inequality.

SDG 5, Gender Equality, aims to achieve gender equality and empower all women and girls. Gender equality intersects with all the SDGs and is therefore essential to advancing sustainable development globally.

We invited trans women, non-binary people, and cis women affiliated with Georgia Tech (undergraduate and graduate students, staff, faculty, researchers, and artists) to submit digitized photography, paintings, creative writing, and research papers, for the exhibition, *Gender Equality: Reimagining our Future through Art and Technology*.

The exhibition uses Gender Equality as a theme to connect diverse research methods, artistic endeavors, and knowledge production occurring today on Georgia Tech's campus. It is not a space to simply showcase women in technology, but to demonstrate how women in technology are reshaping research questions and pushing artistic boundaries which can bring us closer to accomplishing this grand goal.

This is a unique opportunity for the Georgia Tech community to come together from disciplinary perspectives to be inspired by creative merits occurring on campus to reflect, connect, and reimagine the future of Georgia Tech.

This digital magazine includes images of posters and banners on display in The Kendeda Building. It also features research papers, biographies of the artists, and abstracts of their selected work.

SUSTAINABLE DEVELOPMENT GOALS

1 NO POVERTY

2 ZERO HUNGER

3 GOOD HEALTH AND WELL BEING

4 QUALITY EDUCATION

5 GENDER EQUALITY

6 CLEAN WATER AND SANITATION

7 AFFORDABLE AND CLEAN ENERGY

8 DECENT WORK AND ECONOMIC GROWTH

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

10 REDUCED INEQUALITIES

11 SUSTAINABLE CITIES AND COMMUNITIES

12 RESPONSIBLE CONSUMPTION AND PRODUCTION

13 CLIMATE ACTION

14 LIFE BELOW WATER

15 LIFE ON LAND

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

17 PARTNERSHIPS FOR THE GOALS

SUSTAINABLE DEVELOPMENT GOALS

Exhibition Support

The exhibition is supported by The Center for Serve-Learn-Sustain (SLS), Women's Resource Center (WRC), and The Kendeda Building For Innovative Sustainable Living (The Kendeda Building).

SLS is a campus-wide academic initiative working with all six colleges to equip Georgia Tech students to learn and serve around the theme "creating sustainable communities" through engagement with content and context.

WRC advances gender equity across identities by cultivating opportunities for community building, transformative learning, collaborative leadership, and identity development for graduate and undergraduate women.

The Kendeda Building is the latest example of the Georgia Institute of Technology's sustainability leadership and innovation. Georgia Tech occupied the building in September 2019 and constructed it to the Living Building Challenge 3.1 ("LBC") certification standard, the world's most ambitious building performance standard.

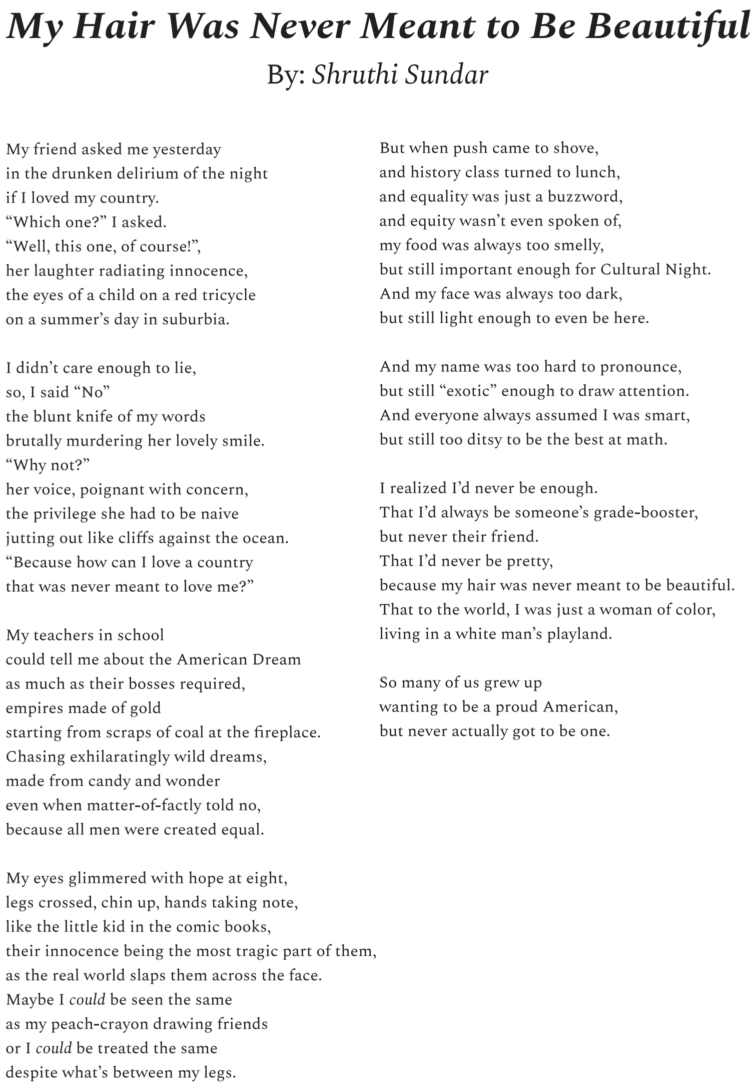
Michelle Ramirez (she/her) is a second-year Digital Media master student and current Graduate Research Assistant with SLS. Under the supervision of Dr. Rebecca Watts Hull (she/her), Michelle researches how to integrate Sustainable Development Goals into the university curriculum. Michelle is responsible for organizing the exhibition.

Contents

Shruthi Sundar

Shruthi Sundar is a BS/MS student in Computer Science who just started her first semester of MS. She is concentrating in Human-Computer Interaction, and is passionate about the use of technology in a social impact sphere, specifically with the Computer Science education equity gap in schools. She has always been passionate about addressing women's rights and has found writing to be her outlet for it. Outside of these, she enjoys skateboarding, climbing, watching movies, and finding street art.

My Hair Was Never Meant to Be Beautiful



My Hair Was Never Meant to Be Beautiful

By: *Shruthi Sundar*

My friend asked me yesterday
in the drunken delirium of the night
if I loved my country.
"Which one?" I asked.
"Well, this one, of course!",
her laughter radiating innocence,
the eyes of a child on a red tricycle
on a summer's day in suburbia.

I didn't care enough to lie,
so, I said "No"
the blunt knife of my words
brutally murdering her lovely smile.
"Why not?"
her voice, poignant with concern,
the privilege she had to be naive
jutting out like cliffs against the ocean.
"Because how can I love a country
that was never meant to love me?"

My teachers in school
could tell me about the American Dream
as much as their bosses required,
empires made of gold
starting from scraps of coal at the fireplace.
Chasing exhilaratingly wild dreams,
made from candy and wonder
even when matter-of-factly told no,
because all men were created equal.

My eyes glimmered with hope at eight,
legs crossed, chin up, hands taking note,
like the little kid in the comic books,
their innocence being the most tragic part of them,
as the real world slaps them across the face.
Maybe I *could* be seen the same
as my peach-crayon drawing friends
or I *could* be treated the same
despite what's between my legs.

But when push came to shove,
and history class turned to lunch,
and equality was just a buzzword,
and equity wasn't even spoken of,
my food was always too smelly,
but still important enough for Cultural Night.
And my face was always too dark,
but still light enough to even be here.

And my name was too hard to pronounce,
but still "exotic" enough to draw attention.
And everyone always assumed I was smart,
but still too ditsy to be the best at math.

I realized I'd never be enough.
That I'd always be someone's grade-booster,
but never their friend.
That I'd never be pretty,
because my hair was never meant to be beautiful.
That to the world, I was just a woman of color,
living in a white man's playland.

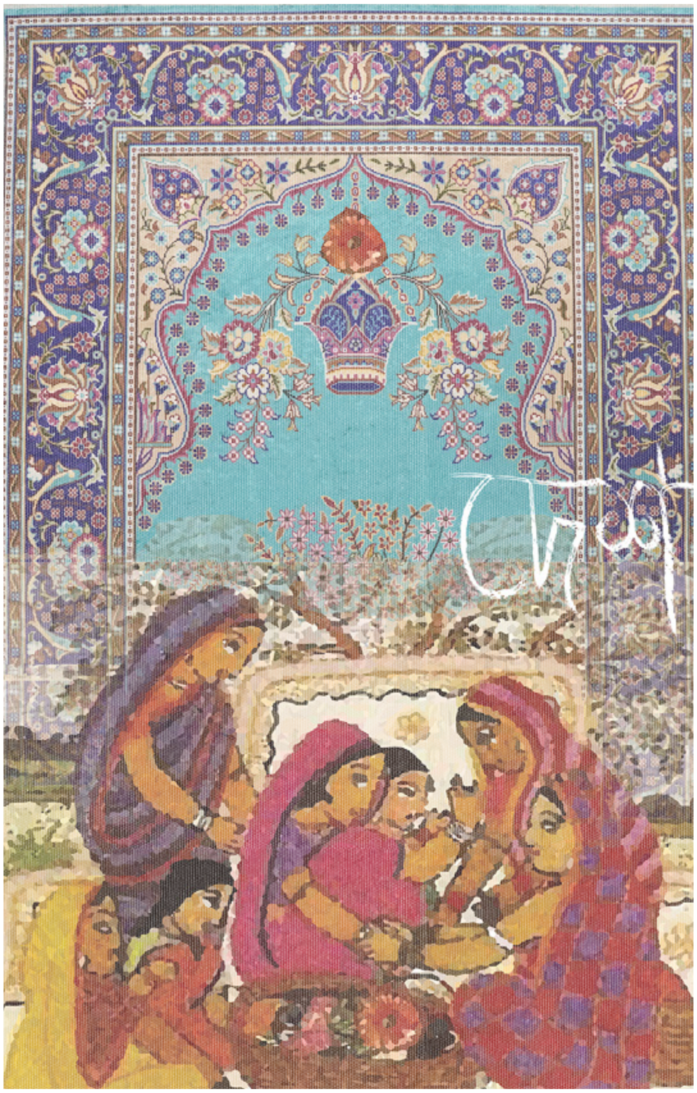
So many of us grew up
wanting to be a proud American,
but never actually got to be one.

My Hair Was Never Meant to Be Beautiful is a coming-of-age piece that addresses the intersectionality of growing up as a woman of color in the US. It touches upon the ways the "American Dream" is shoved down students' throats growing up, despite that dream really only fitting a privileged white male in the US, and how the narrator's childhood innocence takes that dream as her own, despite both fundamentally racist and sexist systems performing against her. This is then juxtaposed against the number of "expectations" set upon girls of color growing up here, including Eurocentric beauty standards, expectations for academic performance in comparison to male peers, and social expectations fitting in with white female peers. This shatters the current notion of the "American Dream" ever being able to fit a growing woman of color, and ends with the realization that the narrator can never call herself a "proud American" because of her identities never being able to fit the mold of "success" she was trained on.

Purna Pratiti Saha

Purna Saha is a second-year Bangladeshi undergrad studying Industrial and Systems Engineering, and Industrial Design at Tech. She loves exploring cultures as a free-flowing stream, and seeks ways of growth thereby. As a Resident Assistant on campus, she gets to share this interest with GT students through hosting events like photowalks, building playlists and trivia on multicultural music.

Feminism in the Womb of Bengal



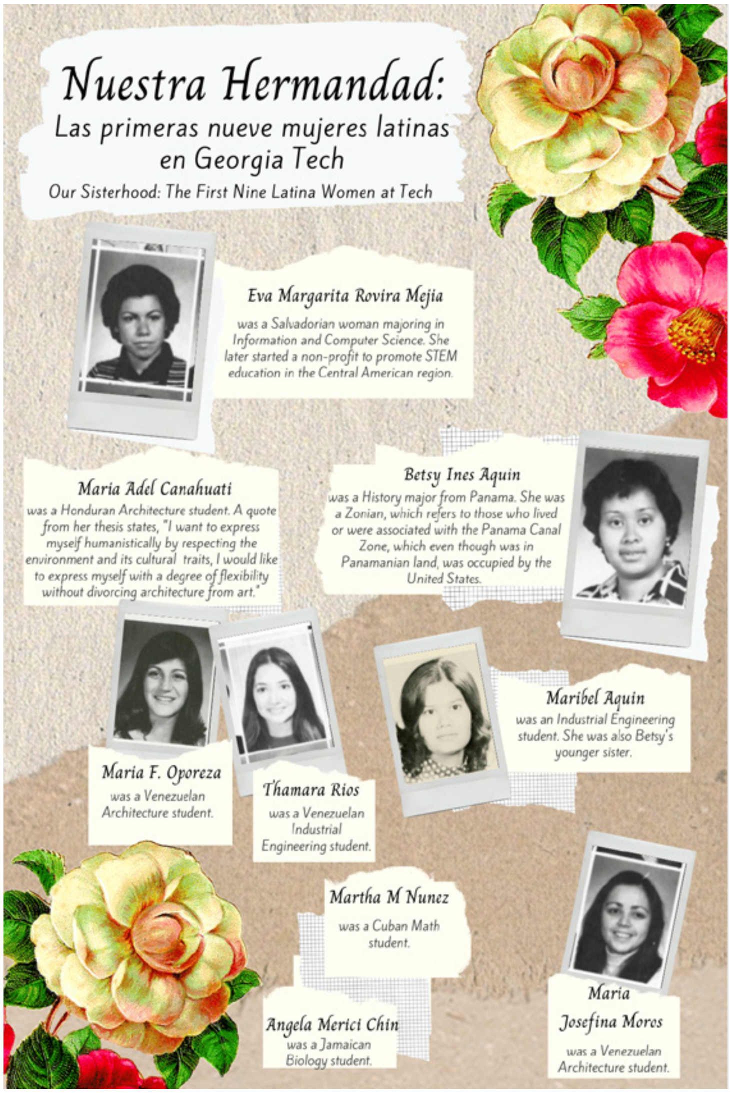
Feminism in the Womb of Bengal is a multimodal essay focusing on the connection between women's empowerment and the city, through a flashback of the cities in Bengal during the modernist era and its reflection in present times. Here I combined two handicraft items to symbolize the gist of the scholarly article "Changing Together, Changing Apart: Urban Muslim and Hindu Women in Pre-Partition Bengal" where the oral accounts of four women—two Hindus and two Muslims—of the first half of the twentieth century were analyzed. I combined two artifacts: a jaynamaz, an Islamic prayer mat, and an embroidered quilt (nakshi kantha) featuring women preparing for a Hindu puja. They seem to blend into each other (although with a distinct partition) by dint of modern education and nationalism—which I symbolize with a chalk-written word— দেশ —desh" meaning "country", a concept gaining high significance as the fight against the colonial rule climaxed in the first half of the twentieth century. The full essay can be viewed at tinyurl.com/bengalwp.

Alexandra Rodriguez Dalmau

Alexandra Rodriguez Dalmau was born and raised in Santo Domingo, Dominican Republic. She is now an undergraduate, studying environmental engineering at Georgia Tech. Before she moved to the U.S., she started a non-profit in the DR called "STEM para Nosotras" with the purpose of providing girls a safe space to develop an interest in the STEM areas as well as motivating them to pursue higher education. Currently, Alex is doing an independent study to learn more about informality, politics, and climate change in the DR.

Nuestra Hermandad: Las primeras nueve mujeres latinas en Tech

is originally an English 1102 class project in which the goal was to use the Georgia Tech archives and research skills to answer a question about Georgia Tech history. I wondered about the first girl from Latin America (Latam) and their story. By requesting some information from the library archives, I was able to find a list of the international students at Georgia Tech since the 60s. In addition, I found other interesting information such as "International Students information packets" and old handwritten calculations for the number of spanish-speaking students on campus. At first, as expected, there were only men international students from Latam. I kept reading and passing the year waiting for a woman to come up. At some point, nine women appeared. After that, I started researching online for any and all information I could get my hands on. In the end, I wrote a narrative so I could share their lives and stories. Read the full blog here.



Nuestra Hermandad:
Las primeras nueve mujeres latinas en Georgia Tech
Our Sisterhood: The First Nine Latina Women at Tech

Eva Margarita Rovira Mejia
was a Salvadorian woman majoring in Information and Computer Science. She later started a non-profit to promote STEM education in the Central American region.

Maria Adel Canahuati
was a Honduran Architecture student. A quote from her thesis states, "I want to express myself humanistically by respecting the environment and its cultural traits, I would like to express myself with a degree of flexibility without divorcing architecture from art."

Betsy Ines Aquin
was a History major from Panama. She was a Zonian, which refers to those who lived or were associated with the Panama Canal Zone, which even though was in Panamanian land, was occupied by the United States.

Maribel Aquin
was an Industrial Engineering student. She was also Betsy's younger sister.

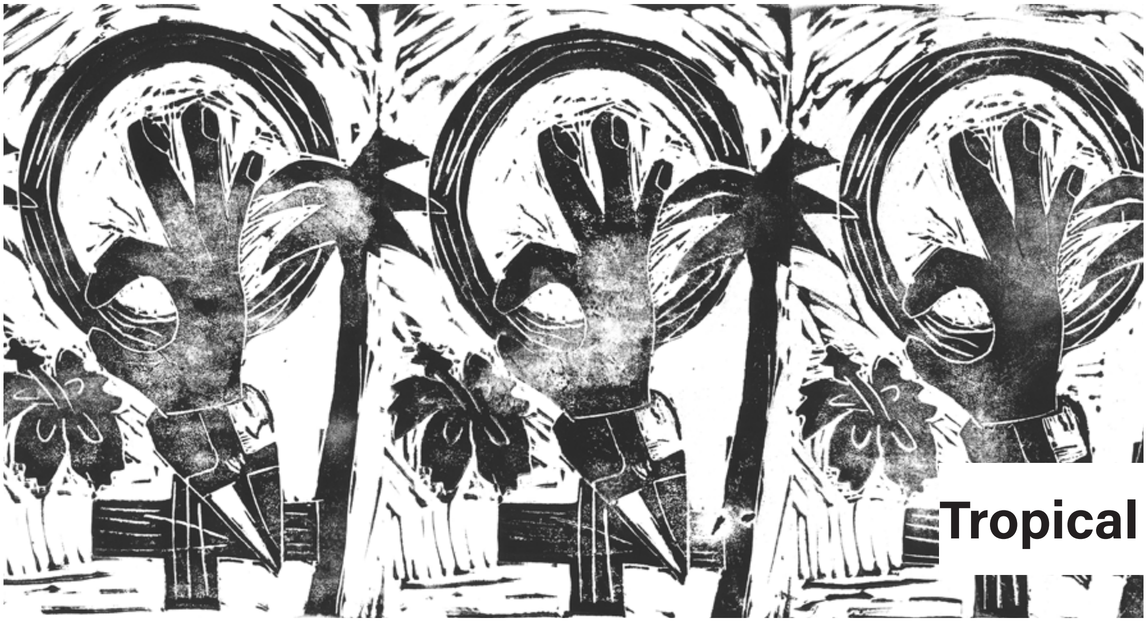
Maria F. Oporeza
was a Venezuelan Architecture student.

Thamara Rios
was a Venezuelan Industrial Engineering student.

Martha M Nunez
was a Cuban Math student.

Angela Merici Chin
was a Jamaican Biology student.

Maria Josefina Moros
was a Venezuelan Architecture student.

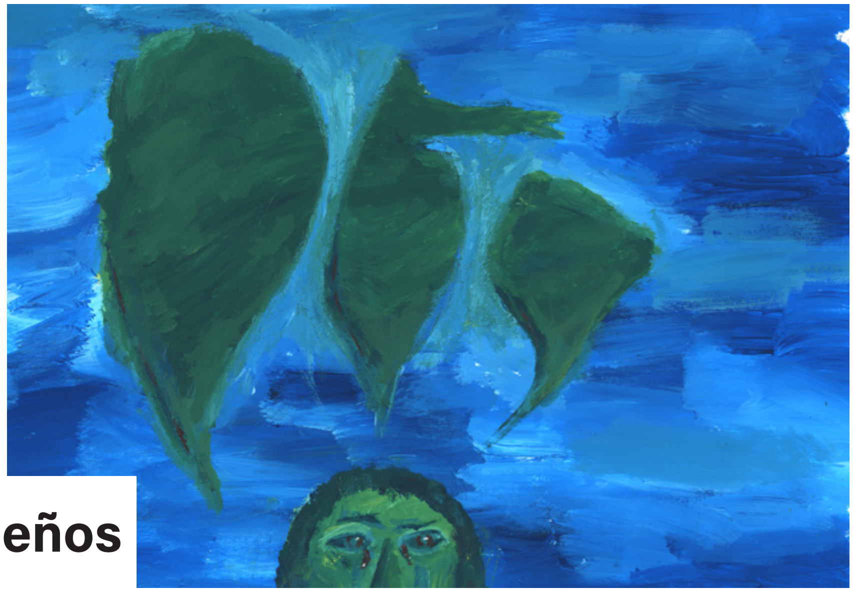


Tropical Feminism

The Dominican Republic is ranked highly among the nations most vulnerable to climate change. My city, Santo Domingo, is going to be one of the most affected by sea-level rise by 2050. Since girls and women are often the most affected by the world's pressing problems, particularly those related to poverty and climate change, it seems only natural to me that incorporating them into the design of their solutions would yield better results. As I'm doing research about my country, I can't help but feel it's too late. I feel nostalgic for a place I haven't lost yet. I feel powerless against such an imminent threat. In the painting, Pensamientos Isleños, it shows the Dominican Republic divided and drawn as thoughts, separated or as if dissolving in water.

Although viewed as a paradise on earth for many who visit, the Dominican people, especially women, do not feel as lucky as tourists do. The Dominican Republic is one of the few countries of the world to have a complete ban on abortion with no exceptions, even when a woman's life is at risk.

Dominican feminists have an ongoing fight to include the "3 causales" aka three circumstances in which they believe abortion should be decriminalized. In the block print art piece, Tropical Feminism, you can see a hand with three fingers lifted, often seen at abortion prostests. The hand appears in front of the female sign and has a handkerchief with the dominican flag. Along the sides appears a palm tree and a "cayena" flower, both of which are often used as tropical and touristic symbols.



Pensamientos Isleños

Katherine E. Bennett

Katherine is a PhD student in Digital Media. Their work probes intersections of race, gender, technology, and environmental justice.

Letters & Editors

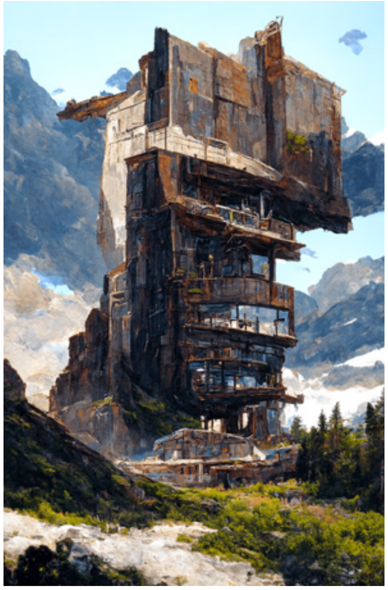
Letters & Editors is a digital craft project that considers historic interactions between race, gender, and place. The project's "letter" interfaces combine text, textile, and computational media to ask how these media can register social-environmental relationships specific to transitioning agricultural production in Jim Crow-era United States. My collaborative reproductions of letters between Georgia farmer Asa Bennett and Pulitzer-winning newspaper editor Ralph McGill make these relationships explicit in text extracted from the Bennett-McGill correspondence and an editorial published in the Atlanta Constitution (1944). My collaborative re-editing of Bennett's raced and gendered letters excavates a message of environmental justice grounded in the labor of reviving a former cotton plantation.

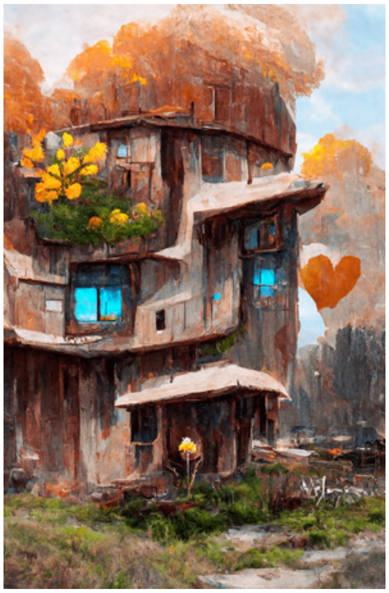
Inspired by craft-based designs (Noel 2020), Letters & Editors combines analog and digital technologies of weaving, sewing, sensing, and coding toward a material dialogue (Nitsche & Zheng 2018) with cotton and rayon—a forestry product that replaced cotton in Southern textile mills during World War II. Nonhuman editors first revised the hybrid letter-objects, buried for three months among the communicative root networks of trees and mushrooms at the Bennett farm. Embedded sensors recorded soil hydration and temperatures altered by climate change and brutal to forest communities. The farm's forester and a local textile artist-farmer next revised the exhumed and marked-up letters, continuing the dialogue with materials and tools of today's practices. Connecting speculative Black Feminist and Queer Media Studies (Everett 2002; Morris 2016, 2014) with Science and Technology Studies (Tsing 2015), the project traces ways that media both perpetuate and change social-environmental relations.

Dr. Anne Sullivan

Dr. Anne Sullivan is an Assistant Professor of Digital Media and head of the StoryCraft Lab at Georgia Tech. Her research focuses on playful and storied interactive experiences from a feminist and humanistic perspective, with an emphasis on human-centered artificial intelligence (AI). She also studies craft as an analog counterpart to playful and storied interactive experiences, researching in the exciting and emerging field of computational craft. Dr. Sullivan is an award-winning quilter and the concept designer and producer of Loominary – a digital game system controlled with a loom - which has been shown internationally, including at the SAAM Arcade exhibit at the Smithsonian American Art Museum.

Gender Representation in the Post-Anthropocene World





Technology and artificial intelligence (AI) are an intrinsic part of life for many people, including those living in the United States. From Google and Netflix to criminal justice and healthcare, AI is used to make decisions from what movie to watch next to whether someone should be incarcerated. AI systems are often considered impartial, but they are programmed by people (in all our flawed glory) and rely on data that reflect existing systemic biases.

This piece shows artwork generated from a “text to image” AI system called Disco Diffusion. The text to generate the images uses positive attributes that have been previously identified as feminine and masculine coded, particularly when describing jobs or job candidates. The resulting images reflect the aesthetic representations of gender inherent in the AI system and data set.

The two prompts used to generate two images each were:

A beautiful painting of an Adventurous Ambitious Confident Courageous Independent Self-sufficient building in a future landscape, featured on artstation.

A beautiful painting of a Cheerful Gentle Empathetic Nurturing Sensitive Warm building in a future landscape, featured on artstation.

Note: “featured on artstation” is used to define an art style consistent with detailed sci-fi/ fantasy concept art.

“We found no violation!” Twitter’s Violent Threats Policy and Toxicity in Online Discourse

Pooja Casula
Georgia Institute of Technology
pcasula3@gatech.edu

Aditya Anupam
Georgia Institute of Technology
aanupam3@gatech.edu

Nassim Parvin
Georgia Institute of Technology
nassim@gatech.edu

ABSTRACT

Threat moderation on social media has been subject to much public debate and criticism, especially for its broadly permissive approach. In this paper, we focus on Twitter’s Violent Threats policy, highlighting its shortcomings by comparing it to linguistic and legal threat assessment frameworks. Specifically, we foreground the importance of accounting for the lived experiences of harassment—how people perceive and react to a tweet—a measure largely disregarded by Twitter’s Violent Threats policy but a core part of linguistic and legal threat assessment frameworks. To illustrate this, we examine three tweets by drawing upon these frameworks. These tweets showcase the racist, sexist, and abusive language used in threats towards those who have been marginalized. Through our analysis, we highlight how content moderation policies, despite their stated goal of promoting free speech, in effect, work to inhibit it by fostering an online toxic environment that precipitates self-censorship in f ear of violence and retaliation. In doing so, we make a case for technology designers and policy makers working in the sphere of content moderation to craft approaches that incorporate the various nuanced dimensions of threat assessment toward a more inclusive and open environment for online discourse. CONTENT WARNING: This paper contains strong and violent language. Language Analysis, Marginalization

CCS CONCEPTS

• Human-centered computing → Collaborative and social computing; Collaborative and social computing theory, concepts and paradigms; Social media; • Social and professional topics →Computing/technology policy; Censorship; Hate Speech.

KEYWORDS

Violent Threats, Online Moderation, Social Media, Free Speech, Twitter, Online Toxicity, Language Analysis, Marginalization

ACM Reference Format:
Pooja Casula, Aditya Anupam, and Nassim Parvin. 2021. “We found no violation!”: Twitter’s Violent Threats Policy and Toxicity in Online Discourse. In C&T ’21: Proceedings of the 10th International Conference on Communities & Technologies - Wicked Problems in the Age of Tech (C&T ’21),

ACM ISBN 978-1-4503-9056-9/21/06. . . $15.00
https://doi.org/10.1145/3461564.346158
June 20–25, 2021, Seattle, WA, USA. ACM, New York, NY, USA, 9 pages.
https://doi.org/10.1145/3461564.3461589

1 INTRODUCTION

In 2018, U.S. political analyst and commentator Rochelle Ritchie reported threatening tweets aimed at her, made by an individual named Cesar Sayoc, on Twitter [1]. These tweets, however, were deemed ‘non-threatening’ by Twitter’s Violent Threats policy. Ritchie’s report was dismissed, and the tweets remained on the platform. Two weeks later, Cesar Sayoc was arrested for sending pipe bombs to prominent U.S. political leaders, including former President Barack Obama. Twitter apologized to Rochelle Ritchie, stating that the company had made an egregious error in dismissing her report [2, 3]

Rochelle Ritchie's experience with Twitter's limited moderation is, unfortunately, not an anomaly. The company's rationale for this minimal approach to moderation is justified by their goal, "to make Twitter a safe place for free expression" [4], despite many Twitter users' contrary arguments and evidence about how this seemingly inclusive moderation policy has a silencing effect [5]. More specifically, the consequences of Twitter's Violent Threats policy are twofold: one, it negatively impacts how users express themselves or self-censor on the platform, and two, it is commonly referred to as the reason why they leave the platform or decide not to join in the first place [6, 7]. So, we might ask: In what ways does Twitter's Violent Threats policy fall short in identifying and removing violent threats on its platform and what are its consequences?

This paper critically examines Twitter's approach to threats of violence by comparing it to linguistic and legal frameworks of threat assessment. Linguistic analyses of violent and abusive speech have played a major role in developing methods of threat assessment and are key to understanding how Twitter's moderation policy dismisses the presence of linguistic indicators of potential violence. We also draw upon legal frameworks that further build on linguistic approaches, to highlight how the policy is dismissive of threatening speech that could even be subject to prosecution in a U.S. court of law. Together, these two approaches open this research space up to an extensive body of work that helps foreground Twitter's shortcomings as a public social media platform with broader implications for public and civic discourse.

2 BACKGROUND

Much of the research on Twitter's content moderation has been centered on hate speech identification and automatic removal. Many studies on hate speech investigate the forms and types of abusive speech that are prevalent on Twitter, their dominant characteristics, as well as the demographics of those who are often the target of such tweets [8–10]. For example, one study found that 89% of , the word was being reclaimed by two women who were usithe hate speech tweets collected on Twitter targeted individuals based on race, behavioral characteristics ("insecure people", "sensitive people"), and physical traits ("short people", "obese people") [9].

Another study investigating rape threats on Twitter applied language analysis techniques to tweets to understand the rhetoric surrounding sexual aggression used in online discourse [10]. Such studies have worked to develop an understanding of how online abusive speech is constructed and who are the primary targets of it. Another area of research in the sphere of content moderation has centered on developing more effective ways of identifying abusive and threatening speech for automatic detection and removal. Many studies focus on how machine learning techniques can be used to conduct automated speech detection [11–13]. The techniques developed in these studies build upon linguistic research for the identification of structure and form of threats to inform automatic classification and identification methods.

These approaches, however, are not without their shortcomings as many scholars have argued. Limitations of algorithmic abusive speech moderation on social media include the inherited bias of machine learning training data and the lack of contextual awareness—both of which disproportionately impact those who have been marginalized. More specifically, algorithmic detection of hate speech both fails to detect toxic speech targeting communities that have been marginalized and blocks non-toxic speech made by members of those communities [14, 15]. For instance, in 2019, Twitter was criticized for blocking a post made by a woman jokingly calling her friend a 'slut'. In a specific context, the word 'slut' is a derogatory slur used against women. However, in this contextng the term to jokingly refer to themselves. Twitter's reliance on algorithms to moderate hateful and abusive content was cited as the reason for the obvious lack of attentiveness to context [16].

In response to such criticism, researchers have developed automated systems to flag abusive speech while simultaneously accounting for context. One study suggested quarantining and reviewing tweets that were flagged by automated systems as being potentially abusive. Such a method would ensure that a human reviewer would look over the quarantined tweets and account for the context in which they were made before removing them or making them public [17]. However, the participation of human moderators is only part of the solution given both the ambiguity of policies and the difficulty of tracing the context and relational specificities of conversations online. Indeed, technology policies and guidelines that are used for threat moderation themselves are not without their own flaws. Many studies have shown that reported complaints by those who were threatened on social media are often ignored as they do not exhibit evidence of violating platform-specific threat moderation policies [18–20].

Twitter's Violent Threats policy specifically has received much media and public criticism for its shortcomings. In 2019, Chad Loder, CEO of the cybersecurity company Habitu8, tweeted several screenshots of death threats aimed at U.S. Congresswoman Ilhan Omar that he found doing a simple search on Twitter. Loder criticized the company's inability to remove clear threats targeting the Congresswoman, highlighting the shortcomings of their moderation system to handle the removal of the widespread toxicity that occurs on the platform [21]. In 2020, Omar herself, along with fellow Congresswomen Alexandria Ocasio-Cortez, Ayanna Pressley, and Rashida Tlaib, openly critiqued Twitter's moderation double standard as manifest in their swift action in removing threats made against then-President Donald Trump, while ignoring death threats made against women politicians [22]. Twitter's minimal violent threat moderation has also been critiqued outside of politics. In a tweet thread created by the Director of Cybersecurity of the Electronic Frontier Foundation Eva Galperin, hundreds of Twitter users detailed their experience receiving threats of violence on the platform and the company's inability to remove them [23].

As noted earlier, the failure to address online hate speech negatively impacts the quality and character of online expression. More importantly, inadequate online threat moderation can in effect restrain or impede freedom of expression beyond online spaces. For instance, in 2019, eighteen female members of the UK Parliament decided to not stand for re-election, citing the inordinate amounts of abuse and threats they received online, and the failure to mitigate them, as one of the main reasons [24]. In another case, former U.S. Congressional Candidate Kim Weaver cited that the main reason she ended her campaign for office was the intense amount of online death threats she received during her campaign [25]. Improvements in threat moderation would help alleviate what has been considered a culture of toxicity prevalent on social media broadly, Twitter, specifically.

Research on threat assessment has not directly engaged with the details of Twitter's Violent Threats policy and how its shortcomings can potentially exacerbate online toxicity. We address this gap, in part, by looking into ways that Twitter's Violent Threats policy falls short and what the resulting consequences are. More specifically, we compare and contrast Twitter's threat assessment policy to linguistic and legal methods of identifying threats to highlight the specific limitations in Twitter's threat moderation. We further analyze three tweets drawing upon the linguistic and legal threat identification frameworks to illustrate how Twitter may be turning a blind eye to users who threaten on the platform. In doing so, we aim to present a case for technology designers and policy makers to engage legal and linguistic scholarship when approaching the problem of content moderation more broadly. We conclude by suggesting further directions for analyzing and moderating social media discourses.

3 SOCIOLINGUISTIC, LEGAL, AND TWITTER'S ANALYSIS OF THREATS

3.1 Linguistic Approach

3.1.1 Linguistic Definitions. The definition of 'threat' has long been a point of philosophic, linguistic, and legal debate. Linguists have generally agreed that threats are "a communication of an intent to harm" [26, 27]. More broadly, drawing upon the work of language philosopher J.L. Austin, a threat can be considered a 'speech act'. Austin argues that there are three acts one could perform when speaking: locutionary acts, perlocutionary acts, and illocutionary acts [28]. Locutionary acts refer to the physical act of speaking.

Perlocutionary acts refer to the effect of speech on the audience. Illocutionary acts refer to the act of intention on behalf of the speaker, or as language philosopher Rae Langton elaborates, “the action performed simply in saying something” [29]. Many linguists have argued that the speaker’s intention, or the illocutionary aspect of speech, is the most important aspect when identifying a threat [27].

However, the emphasis on intent alone has been debated, with scholars arguing that the speaker’s intention, or lack thereof, does not determine whether the statement is considered threatening or not. Linguist Kate Storey argues that the context in which a statement is made, and the interpretation of the hearer are necessary when determining whether the said statement is a threat or not [30]. For instance, the phrase “I’m going to find you. . .” in the context of a game will be interpreted by the hearer as non-threatening. Yet the same utterance would take on a threatening interpretation if the hearer was being stalked. In both cases, the speaker intends to find the hearer. However, in the latter case, it is presumed that the speaker wants to find the hearer with the further intention of harming them. Here, the context of the statement and the interpretation of the hearer are key aspects in deeming the statement to be a threat.

Sociolinguists have identified three main categories of threats based on their content: direct, conditional, and indirect [31]. Direct threats are specific and typically contain phrases that imply the explicit intent to harm. For example, threats with phrases such as “I will [...]” or “I am planning to [...]” are considered direct due to the use of first-person pronouns and decisive verbs such as ‘will’ that imply definitive intent [32]. Like direct threats, conditional threats are also specific and contain explicit intent. They follow an ‘if-then’ format, presenting a condition to the target of the threat directly [33, 34]. An example of a conditional threat would be, “If you don’t stop talking right now, I will kill you”. Indirect threats, also known as veiled threats, are not specific and typically do not contain any direct intent. This form of threat is highly dependent on context. For example, in many instances of veiled threats, the speaker and target share some degree of knowledge that is unknown to witnesses. In such a case, the speaker ensures that the statement will only be found threatening by the target [35]. Linguist Roger Shuy exemplifies this in the statement, “How’s David?” To an outside observer, unaware of the context, this statement appears benign. It appears that the speaker simply wants to know how an individual named David is doing. But in a specific context, in which perhaps David is in danger or missing, the statement shifts from being benign to threatening [36].

3.1.2 Application of Linguistic Definitions. Much research in the area of linguists, especially forensic linguistics, has focused on determining linguistic indicators of potential violence in threats. Researchers have generally found that the more specific a threat is, the more likely it will result in violence. For instance, the statement “I will come to your house with a gun on April 2nd and kill you” is considered more threatening due to its specificity than the statement, “Your time is coming. . .”. Common rhetorical features of threats include the use of first-person pronouns (“I”, “my”), secondperson pronouns (“you”, “your”), obscenities, violent verbs (“kill”, “shoot”), and adverbials of time (“now”, “soon”) [32]. A study found specific themes based on language in threats that were found to be indicators of violence [37]. Some of these themes were: hopelessness, violent behavior, fantasies, intimidating claims, weapons deadlines, and racism. When studying and understanding linguistic features of threats, researchers typically focus on direct and conditional threats, as these are most likely to exhibit indicators of individual or collective intent and are therefore the easiest to identify [33].

However, in some instances, conditional threats are confused for warnings. The existence of a conditional clause may serve to warn the recipient of the harmful consequence of not meeting a specified condition. Some sociolinguists have suggested that what distinguishes threats from warnings is when a speaker makes a threat, they control the outcome; yet when a speaker gives a warning, the hearer controls the outcome [39]. Other sociolinguists would argue that in many cases, the line between threats and warnings are purposefully blurred so that those who made the threat can avoid legal consequence [35, 39]. In such cases, it is the interaction of the intent of the speaker, the context of the situation, and the reaction of the recipient that can clarify whether the statement in question is truly a threat or not.

Indirect or veiled threats are often difficult to identify due to their unpredictable context [33]. With no identifiable structure, it is the very nature of veiled threats that make them more dangerous than direct or conditional threats. Speakers may make veiled threats to create plausible deniability, a situation in which they can reasonably claim that since their statement did not include explicit intent, it should not be considered a threat [45]. For example, consider the phrase "How's the leg?". This phrase could be considered threatening if the speaker hurt the hearer's leg, but to a witness, the phrase would be considered benign. In this case, the speaker could argue that their statement does not implicate themselves as a person who wants to harm the hearer because it contains no direct intention. Since the context of the statement is shared only between the speaker and the recipient, the speaker has created plausible deniability because the only person who would find the statement threatening would be the recipient. This notion that technically any statement, in the right context, can be considered threatening, is also supported by J.L Austin's Speech Act Theory. If the speaker's intention to be threatening was received by the hearer, it can be labeled a threat, regardless of wording [26, 28].

3.2 Legal Approach

3.2.1 Legal Definitions. The First Amendment in the U.S. Constitution states that every individual has the right to free speech. Threats of violence are among the few forms of speech that are unprotected by the law [40]. Though the U.S. Supreme Court has yet to take an official stance on what exactly constitutes a threat of violence, they have delineated a set of three criteria, known as the Watts Factors [41], to help distinguish between a threat and a statement protected by the First Amendment. These criteria are very similar to the three components of speech acts outlined by linguists. They include the context of the statement; the conditional nature of the statement (or the intent of the person making the statement); and the reaction of those who hear the statement. The Watts Factors have been influential in guiding the U.S. Courts of Appeals, otherwise known as circuit courts, in their creation of 'True Threat' tests, which have been used as a form of threat identification in court [42]. The U.S. Courts of Appeals comprises a set of 13 appellate courts with each appellate court serving a specific regional district across the United States. The U.S. Courts of Appeals are considered the most powerful courts in the U.S., second only to the Supreme Court, and thus have a widespread influence in setting legal precedent and determining policy [43].

Law enforcement agencies draw upon the work of forensic linguists to develop their own set of criteria for distinguishing various forms of threats. According to the Federal Bureau of Investigation (FBI), the principal domestic law enforcement agency of the United States, threats fall under four categories: direct, indirect, veiled, and conditional [44]. The FBI classifies direct threats as those that contain explicit intent to harm a specific individual. For instance, the statement, "I am going to place a bomb in the school's gym" is deemed as a direct threat due to its explicit mention of how the individual intends to harm a group of people. Indirect threats are "vague, unclear, and ambiguous" and imply that a violent act could potentially occur. An example of this can be seen in the statement, "If I wanted to, I could kill everyone here!" The FBI defines a veiled threat as "one that strongly implies but does not explicitly threaten violence", the danger being that the target of the threat is left to think about what might happen. The statement, "We would be better off without you around anymore." does not contain any explicit intent, nor does it implicate the speaker as one to do any harm, yet the recipient of such a statement is left to feel fearful of what could potentially occur to them. Conditional threats are identified as warnings of violence if a demand is not met. For example, "If you don't pay me one million dollars, I will place a bomb in the school" [44].

3.2.2 Application of Legal Definitions. In a U.S. court of law, threats are often distinguished from statements protected by the First Amendment through the use of 'True Threat' tests. There are many different versions of 'True Threat' tests used among different circuit courts, but almost all require evidence of the speaker purposely and knowingly making the statement which tends to be qualified as general intent. Most courts rely upon 'objective' tests, which, in addition to requiring evidence of general intent, emphasize whether a 'reasonable person' would consider the statement in question threatening or not. These 'reasonable person' tests tend to be complicated by whether the court uses the 'reasonable speaker' test or the 'reasonable listener' test. The 'reasonable speaker' test determines the statement to be a threat if the person making the statement anticipates the recipient to interpret it as a threat [42]. Thus, "knowingly transmitting the threat makes the act criminal" [46]. The 'reasonable listener' test determines the statement to be a threat if any person, knowing the full context of the statement, finds the statement threatening.

Many free speech activists have argued that the 'reasonable person' tests are not enough for one to be prosecuted as they can lead to many people being punished for simply being careless with their words at a specific moment in time [46]. Several courts approach the 'True Threat' test with a more 'subjective' approach, emphasizing whether the speaker truly had the 'specific intent to threaten' or the 'specific intent to carry out the threat' [42]. While the Supreme Court has yet to determine the specificity of intent required for a threat to be deemed a threat, in all forms of the 'True Threat' tests, the interpretation of the statement by the recipient or a 'reasonable person' listening, is accounted for, in addition to the general intent of the speaker.

In 2015, the U.S. Supreme Court supported the circuit courts' 'reasonable speaker' test in a case involving online threats. In Elonis v. United States [47], the defendant threatened to kill his ex-wife on Facebook. He claimed that he did not intend to kill his ex-wife, but rather was posting angsty rap lyrics to reflect how he felt about their separation. This case echoes the previously mentioned instance of plausible deniability. In this case, Elonis posted threatening statements with little to no direct intent as a way to later deny that he had any intention of committing harm. While the Supreme Court ruled in favor of Elonis, due to an error regarding how the case was presented to the jury, they did make a judgment regarding how online threats should be considered. The majority opinion stated that for a statement to be considered a threat, the speaker needs to have the intent of making the threat and know that what they post will be interpreted as a threat. As historian Angus Johnston succinctly summarized in his analysis of the Supreme Court opinion, Supreme Court Justice Samuel Alito, in his concurring opinion in the Elonis v. United States case, stated that statements directly addressed to another individual on the Internet can and will be taken more seriously, specifically because of their context [47]. The very nature of Twitter, and social media in general, encourages sharing one's thoughts to a public audience, not necessarily directly addressing one individual. In this context, users can technically make threats on social media without their targets knowing, by simply not tagging them or not mentioning them [49]. Thus, when a user does make a statement about harming an individual and then mentions that person's account, they are ensuring that the tweet will be seen by the

> "If you make a threat online, and you know the person who receives it will see it as a threat, you're guilty of violating federal law. It doesn't matter if you claim that it's protected speech, or put a smiley face at the end, or point out later that the threat was really just lyrics from an old Beatles song. If you send a threat and you know it'll be interpreted as a threat, you're guilty. Period." [48]

person mentioned. A person will take a threat more seriously if they were specifically intended to see it [47]. For instance, a user who simply tweets about harming a U.S. Congresswoman will have less of an impact than a user who tweets directly at the Congresswoman, because that user is ensuring that she will be notified of the message, implying their intent to threaten.

4 TWITTER'S VIOLENT THREATS POLICY

4.1 Twitter's Definition and Application of the Violent Threats policy

Twitter's definition of 'violent threat' differs from the linguistic and legal interpretations. According to the company's Violent Threats policy, only tweets with the stated "intention to inflict violence on a specific person or group of people" are considered threats and removed from the platform [38]. For example, "I will kill you" would be considered a direct threat since it includes explicit intent to harm. In contrast, tweets that make "vague or indirect threats", are excluded and not actionable under their policy [38]. An example of a tweet containing a vague or indirect threat is, "Your time is coming...". Since the statement has no explicit intent and simply hints at potential harm, it is considered non-threatening by Twitter. Put simply, statements that imply a hypothetical or do not carry a specific degree of certainty, whatever that may be, are not classified as threats by Twitter moderators.

In addition to intent, the policy acknowledges the role of context in the decision to classify a statement as a threat. It states: "We recognize that some people use violent language as part of hyperbolic speech or between friends, so we also allow some forms of violent speech where it's clear that there is no abusive or violent intent". For instance, the statement "I will kill you for sending me spoilers!", made between friends, will not be considered a threat under the policy [38]. The sole focus on context and intent, without the inclusion of the lived experiences of harassment by the recipients of such threats, such as their reaction and interpretation, results in the persistence of many tweets that users legitimately find threatening. The problematic nature of this approach is evident when we consider cases such as that of Rochelle Ritchie, as described in the opening of this paper. Her report of Cesar Sayoc's threatening tweet was dismissed, resulting in serious consequences [2]. In other words, compared to linguistic and legal frameworks of threat assessment, Twitter's Violent Threats policy is quite limited. In the following section, we illustrate the relevance and importance of these frameworks for understanding the shortcomings of Twitter's online threat moderation policy through the examination of three tweets. For this purpose, we have selected three tweets aimed at public figures in the U.S.

Of the three tweets we selected, two were aimed at U.S. Congresswoman Alexandria Ocasio-Cortez and one was aimed at U.S. Congresswoman Ilhan Omar. We chose to find tweets targeting these two Congresswomen for two reasons. First, both Congresswoman Ocasio-Cortez and Congresswoman Omar have a large presence on social media, particularly Twitter [50]. Second, both Congresswomen have publicly discussed the inordinate amount of hateful and abusive speech they receive on Twitter specifically [21, 22]. A study conducted by the Institute of Strategic Dialogue found that among the Congresspeople they studied running for reelection in the 2020 U.S. Congressional elections, Congresswoman Omar and Congresswoman Ocasio-Cortez received the highest amounts of online abuse [51].

The rationale for selecting these tweets is threefold. First, they contain a graphic or specific depiction of violence. Second, all three tweets are directly aimed at the Congresswoman in question since they tagged their account. The first and third tweets were made as a direct reply to the Congresswoman's tweet and as a result, directly tagged her account, ensuring that she would see it. The second tweet, while not a direct reply, also directly tagged the Congresswoman's account. Finally, all three tweets remained publicly available on Twitter for an extended time. Two of the three tweets identified remain on Twitter as of April 26th, 2021. The account that made the third tweet was suspended two months after the tweet was made. As a result, the tweet in question is no longer online. We could not determine whether the tweet had any bearing on the suspension of the account.

Each tweet highlights a different form of threat. The first tweet is an instance of a veiled threat of violence. The second tweet is an instance of a conditional threat with mention of explicit violent behavior, and the third tweet is an instance of a threat of sexual violence. Together, these three tweets are illustrative of the toxic nature of discourse on social media, particularly threats and abuse targeting women and people of color. They highlight how such threats can escape moderation when the broader socio-political, cultural, and historical context of public discourse is not accounted for.

The point of drawing upon linguistic and legal analyses of threatening speech with the following three tweets is not to claim that such a tweet could be contested as a threat of violence in a U.S. court of law. Such a claim is outside the scope of this paper and the authors' expertise. Rather, drawing upon the nuances of other wellestablished bodies of knowledge in identifying threats—such as that of sociolinguistic and legal scholars—could shed light on Twitter's threat moderation limitations. In other words, we draw upon linguistic and legal frameworks to foreground how Twitter's Violent Threats policy fails those who are threatened on the platform.

4.2 Three Illustrative Tweets

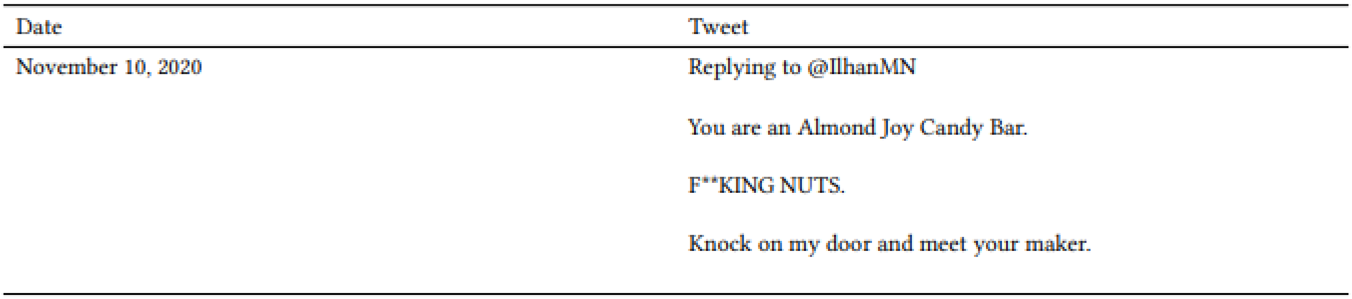
4.2.1 Tweet #1. In Tweet#1 (see Table 1), the user makes a demeaning comment to Congresswoman Ilhan Omar, followed by a veiled reference to killing her: "meet your maker." This tweet contains several linguistic indicators of threatening speech. By directly replying to the Congresswoman, the author of the tweet is directly targeting her. The tweet mentions "you" and "my", words that have been found to be potential indicators of violence as they form a direct connection between the author of the tweet and the recipient of the tweet [37]. It is evident that in this case, the author of the tweet has implicated themselves as an individual who would like for harm to come to the Congresswoman, or more likely, to harm the Congresswoman themselves.

This tweet could also be deemed threatening by legal standards as it could pass a few of the 'True Threat' tests used in circuit courts. While it is difficult to know whether the user had the intention of acting upon their statement or how Congresswoman Omar interpreted the statement, the tweet does contain both general intent and the specific intent to threaten according to the legal standard of the courts. The tweet includes general intent as the user made the deliberate choice to tweet the statement. The tweet also includes the specific intent to threaten, indicated by the fact that the tweet was in direct response to Congresswoman Omar's tweet. This implies that the user ensured that Congresswoman Omar would see the tweet. One may even argue that the general intent and the evidence for the specific intent to threaten implies that the tweet also passes the 'objective' 'reasonable speaker' test [42]. The speaker stated the circumstances in which the Congresswoman would die and then proceeded to directly mention her Twitter account, implying that they had to be reasonably aware that the recipient would find the statement threatening.

We can only speculate regarding why Twitter moderators consider this tweet to not violate their Violent Threats policy. One reason could be that, though this tweet establishes a direct connection between the individual who tweeted and the recipient, the tweet lacks explicit intent to harm. The tweet does not contain definitive verbs ("will") or violent verbs {"kill"), nor does it contain the use of first-person pronouns ("I"), all common examples of explicit intent. The tweet is quite vague, as the use of the phrase "meet your maker" can be considered an indirect or euphemistic way of saying "to die" [52]. This, coupled with the general absurdity of the tweet may have been considered signs of hyperbolic speech, a form of speech not meant to be taken seriously [53]. However, one cannot dismiss the racist and sexist language of the tweet itself, and how the recipient would reasonably interpret this statement as a threat of harm. This tweet serves as an example of how Twitter may disregard threats that utilize euphemistic language as a mechanism for veiling the violence and therefore evade any consequences under the current platform policies.

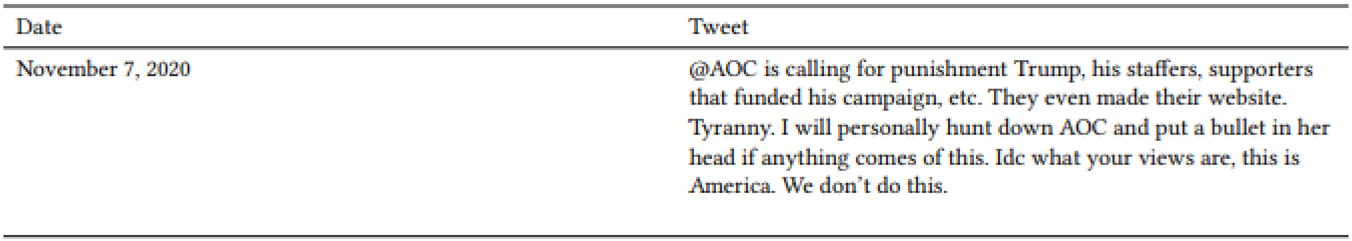
4.2.2 Tweet #2. Tweet#2 (see Table 2) directly tags Congresswoman Ocasio-Cortez's Twitter account and is making a statement referencing a previous tweet made by the Congresswoman. On November 6th, 2020, Congresswoman Ocasio-Cortez made a tweet about archiving tweets and media made by

Table 1: Tweet#1 was made on November 10th, 2020 and as of April 26th, 2021, is still publicly available.



Date	Tweet
November 10, 2020	Replying to @IlhanMN You are an Almond Joy Candy Bar. F**KING NUTS. Knock on my door and meet your maker.

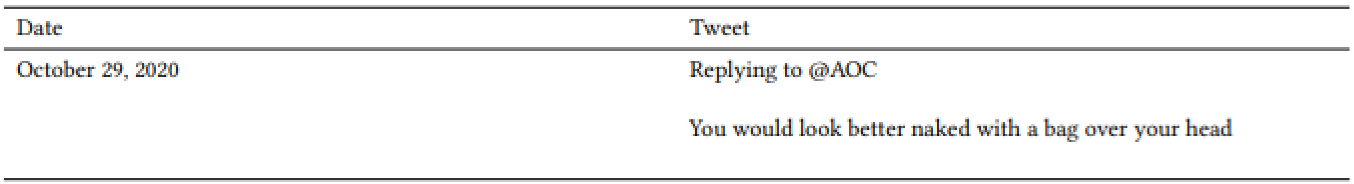
Table 2: Tweet #2 was made on November 7th, 2020 and as of April 26th, 2021, is still publicly available on Twitter.



Date	Tweet
November 7, 2020	@AOC is calling for punishment Trump, his staffers, supporters that funded his campaign, etc. They even made their website. Tyranny. I will personally hunt down AOC and put a bullet in her head if anything comes of this. Idc what your views are, this is America. We don't do this.

then-President Donald Trump's supporters who might 'downplay' their 'complicity' in supporting his Presidency [54]. Following this tweet, the Trump Accountability Project, a group created with the purpose of holding then-President Trump and his supporters accountable for their actions, released a statement made on their website. The statement called for people associated with the Trump administration to, in essence, be blacklisted from future job opportunities [55]. The Trump Accountability Project website was later taken down. This tweet would be characterized as a conditional threat by linguists. Conditional threats typically follow the 'if-then' form, making them quite easy to identify [33, 34]. While this tweet does not strictly follow the 'if-then' format, it does contain a conditional clause when the author of the tweet states, "I will personally hunt down AOC and put a bullet in her head if anything comes of this", "this" referring to people associated with the Trump administration being potentially blacklisted. The tweet also contains other linguistic features of threats such as the use of personal pronouns, the specific mention of a weapon, the use of words such as "will" which signify intent and action, and the use of the violent verb "hunt" which is often used in the context of chasing to capture or harm [27, 31, 56]. From a legal standpoint, it may be argued that such a tweet be classified as a threat of violence as it passes several 'True Threat' tests acknowledged by the circuit courts–even though it is difficult to know the Congresswoman's reaction to the tweet or the user's intention to carry out the threat. The tweet includes a general intent as the author of the tweet made the conscious decision to type and post the tweet. It also includes the specific intent to threaten, as the user directly tags the Congresswoman's account to ensure that she views the statement that they made regarding their intention to "hunt" her and shoot her. While this tweet remains on the platform, it is known that Twitter has flagged tweets with similar statements before, providing evidence that moderators at Twitter consider the mention of a weapon and subsequent violent phrase in this tweet as a threat of violence [57]. Also, regardless of the precedent, the use of the phrase "I will personally hunt down AOC", displays explicit intent, even by Twitter's standard. One can speculate that the reason Twitter has let the tweet remain on the platform is because of its conditional nature. The user claims that if, and seemingly only if, people who are associated with the Trump administration are truly blacklisted, then they will act upon their threat. The 'if' implies a degree of ambiguity and uncertainty and thus moderators may have felt that it need not be acted upon. That being

Table 3: While Tweet#3 remained on Twitter until at least December 30th, 2020, this account has been suspended as of April 26th, 2021. It is unknown for what specific reason this account has been suspended.



Date	Tweet
October 29, 2020	Replying to @AOC You would look better naked with a bag over your head

said, the threat made in the tweet is quite direct and explicit, warranting a truly specific and valid reason from Twitter as to why it has not been taken down yet.

4.2.3 Tweet #3. Tweet#3 (see Table 3) was made in response to Congresswoman Ocasio-Cortez's tweet regarding the reaction of many of her colleagues to her appearance on the cover of the wellknown fashion magazine, Vanity Fair [58]. This tweet is arguably referencing both physical and sexual violence against the Congresswoman.

From a sociolinguistic standpoint, such a tweet could be considered a veiled threat. While the user provides a harmful description, they do not implicate themselves as the person doing any harm [26]. Put differently, since the user does not directly say that they will harm the Congresswoman themself, or use personal pronouns, the tweet lacks indicators of direct intent. The tweet also does not use any explicit violent verbs or profanity, other linguistic indicators of threatening violence [32]. However, this tweet does contain themes that some linguists would argue serve as a measure of potential violence [32, 37]. The tweet references violent sexual behavior directly targets an individual through the use of second-person pronouns ("you, your head"), and contains a specific method of physical violence (suffocation). Though this tweet was not flagged by Twitter moderators for an extended time, it does pass a few 'True Threat' tests and could be considered a threat of violence in a U.S. court of law. The tweet contains general intent as the user made the deliberate decision to type out and post the tweet. The tweet also contains the specific intent to threaten as the author uses the word 'you', directly addressing the Congresswoman, in addition to, directly replying to her tweet. The user ensured that she would be able to view their response.

This tweet also passes the 'reasonable speaker' test, as the user had to have known such a tweet would be interpreted as a threat of sexual violence. When the user uses the phrase, 'You would look better naked. . .', it implies, but arguably does not confirm, that the Congresswoman would look better naked to them specifically. While the user does not use the word "I" or other personal pronouns [31], one can only reasonably assume that the user has made this threat implying that they are the person who intends to attack the Congresswoman and place a bag over her head. Even if this is not the case, the tweet can also serve to instigate others who have viewed this tweet to potentially act on its threat. It is also easy to imagine that anyone on the receiving end of such a statement would reasonably interpret it as a threatening remark.

One can speculate that the main reason why this statement did not violate Twitter's Violent Threats policy is that it did not contain explicit intent. Since the user did not implicate themselves as a person doing any harm, and simply hinted at the potential of harm, the tweet technically did not violate the policy. If this is indeed the case, this tweet serves as an example of how, by only looking for intent words and phrases in tweets, Twitter's Violent Threats policy can fail to consider how one may have interpreted such a graphic and explicit statement. In this case, the policy gives a clear pass to users who are able to make an indirect threat by circumventing Twitter's focus on identifying words and phrases featuring direct intent.

Together, the above tweets illustrate that incorporating context, intent, and interpretation when determining a tweet violation would serve as a good first step for Twitter to create a more inclusive, inviting online environment that does not leave a victim of a threat powerless in the face of the abuser.

5 DISCUSSION

The failure of Twitter's threat moderation policy to mitigate threats against powerful politicians is indicative of the broader failure to address violent threats toward public figures and the general public. On the surface, this might seem like an effort to protect free speech, as indicated by the stated goal of the platform [4, 59]. However, free speech is also contingent upon a robust and reliable approach to regulating violent speech if we were to address various forms of self-censorship in fear of violence and retaliation.

That is not to say that content moderation is an easy task. On the contrary, it is a particularly nuanced and difficult practice. The problem of content moderation reflects the broader characteristics of problematic ethical situations, inclusive of the uncertainty that permeates such situations [60]. Linguistic and legal frameworks of threat assessment offer valuable insights into the moderation of online violent threats, especially in their inclusion of how a threat is perceived and experienced, as illustrated in this paper. However, analyzing a tweet for expressions of violence cannot be limited to a focus on linguistic indicators or the application of a set of rules no matter how complex. Rather, content moderation entails accounting for all the different qualities of a rhetorical situation. This includes not only the specific lived experiences and reactions of individuals who are the targets of threats but also the interpersonal, social, political, and historical dynamics and trajectories embodied in the situation. For example, Tweet #3, “You would look better naked with a bag over your head” is undeniably an expression of violence not only because of linguistic indicators but also because it echoes violent histories of misogyny and racism that all too often mark violent threats against women of color. The tweet can be considered not only an isolated threat to an isolated person but also, a threat to free and democratic discourse as it embodies expressions that have historically been used to silence and disparage women and people of color. To overlook the necessity to remove such a tweet under the guise of freedom of speech is to overlook and validate troubling historical trajectories and patterns of behavior complicit in the collective harassment and silencing of voices that have been marginalized.

So, the question remains: whose free speech is Twitter’s Violent Threats policy protecting and whose voices are being silenced in turn? Why is it that the burden of marking, reporting, and documenting tweets of violence is disproportionately placed on the recipients, whose complaints are often disregarded or ignored? Alternatively, we might wonder what it would be like to participate in a platform that is serious about accounting for experiences of violence and silencing– siding with those who are being threatened by mitigating violent and intimidating tweets no matter how veiled or ambiguous? Such an approach would entail having regard for power differentials that mark the relational aspect of the speakers and audiences as well as the socio-political, cultural, and historical context of public discourses and public fora. The task of online moderation is not simply one of identifying and removing abusive and hateful speech. It is a societal problem that is fundamental to how individuals and groups experience and practice freedom of expression without fear.

6 CONCLUSION

Threat moderation policies on social media platforms present many challenges. Twitter’s Violent Threats policy has specifically been criticized by several prominent public figures for its failure to mitigate online toxicity. In this paper, we drew attention to the shortcomings of Twitter’s Violent Threats policy by comparing it to the threat assessment frameworks of both linguistic and legal scholars. By showcasing the similarities and differences of these frameworks, we highlighted the importance of incorporating how one may interpret or perceive a tweet as part of Twitter’s threat moderation policy (as well as other social media platforms). We also draw upon these frameworks to analyze three specific tweets, which, more broadly, illustrate how content moderation policies in their aim to promote free expression, may in effect work to hinder it. Through this process, we present a case for technology designers and policy makers to approach the problem of content moderation through an interdisciplinary perspective. This research also highlights that much remains to be explored within the realm of content moderation on social media including but not limited to how online toxicity influences public and democratic discourse offline; how language and discourses of violence against women and people of color are echoed and amplified in virtual spaces; and mechanisms that might foster more inclusive online environments.

REFERENCES

[1] Jessica Guynn. 2018. Trump, Twitter to blame for Cesar Sayoc threats, Rochelle Ritchie says. USA Today. Retrieved February 25, 2021 from https://www.usatoday.com/story/news/2018/10/26/cesar-sayoc-threatsrochelle-ritchie-blames-twitter-trump/1779398002/

[2] Azmina Dhrodia. 2018. Why is Twitter still not acting when it receives reports of death threats? Amnesty International. Retrieved February 25, 2021 from https://www.amnesty.org/en/latest/news/2018/10/why-is-twitter-still-notacting-on-reports-of-death-threats/

[3] Twitter Safety. 2018. An update. We made a mistake when Rochelle Ritchie first alerted us to the threat made against her. The Tweet clearly violated our rules and should have been removed. We are deeply sorry for that error. @TwitterSafety. Retrieved April 27, 2021 from https://twitter.com/TwitterSafety/status/1055999127446728704

[4] About Twitter | Healthy conversations. Retrieved February 25, 2021 from https://about.twitter.com/en/our-priorities/healthy-conversations.html

[5] 2017. "We found no violation of Twitter's Rules." Medium. Retrieved April 27, 2021from https://medium.com/@no_violation_of_twitter_rules/abuse-98002aaf35d8

[6] Rebecca Macatee. 2014. Jennifer Lawrence Will Never Join Twitter, SocialMedia "Because the Internet Has Scorned Me So Much." NECN. Retrieved April 29, 2021 from https://www.necn.com/news/national-international/jenniferlawrence-will-never-join-twitter-social-media-mockingjay/2040829/

[7] Lindy West. 2017. I've left Twitter. It is unusable for anyone but trolls,robots and dictators | Lindy West. the Guardian. Retrieved April 27, 2021
from http://www.theguardian.com/commentisfree/2017/jan/03/ive-left-twitterunusable-anyone-but-trolls-robots-dictators-lindy-west

[8] Hassan Mohamed, Syahaneim Marzukhi, Zuraini Zainol, Tengku Mohd Tengku Sembok, and Omar Zakaria. 2018. Semantic-based Social Media Threats Detection. In Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication (IMCOM '18), Association for Computing Machinery, New York, NY, USA, 1–4. DOI:https://doi.org/10.1145/3164541.3164620

[9] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In Proceedings of the 28th ACM Conference on Hypertext and Social Media (HT '17), Association for Computing Machinery, New York, NY, USA, 85–94. DOI:https://doi.org/10.1145/3078714.3078723

[10] Claire Hardaker and Mark McGlashan. 2016. "Real men don't hate women":Twitter rape threats and group identity. Journal of Pragmatics 91, (January 2016),
80–93. DOI:https://doi.org/10.1016/j.pragma.2015.11.005

[11] Noman Ashraf, Rabia Mustafa, Grigori Sidorov, and Alexander Gelbukh. 2020.Individual vs. Group Violent Threats Classification in Online Discussions. In Companion Proceedings of the Web Conference 2020 (WWW '20), Association for Computing Machinery, New York, NY, USA, 629–633. DOI:https://doi.org/10.1145/3366424.3385778

[12] Aksel Wester, Lilja Øvrelid, Erik Velldal, and Hugo Hammer. 2016. Threat detection in online discussions. 66–71. DOI:https://doi.org/10.18653/v1/W16-0413

[13] Katie Cohen, Fredrik Johansson, Lisa Kaati, and Jonas Clausen Mork. 2014. Detecting Linguistic Markers for Radical Violence in Social Media. Terrorism and Political
Violence 26, 1 (January 2014), 246–256. DOI:https://doi.org/10.1080/09546553.2014.849948

[14] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In Social Informatics (Lecture Notes in Computer Science), Springer International Publishing, Cham, 405–415. DOI:https://doi.org/10.1007/978-3-319-67256-4_32

[15] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 7, 1 (January 2020), 2053951719897945. DOI:https://doi.org/10.1177/2053951719897945

[16] Ana Valens. 2021. Twitter suspended me for making a "slut" joke. But what if I am a slut? The Daily Dot. Retrieved April 27, 2021 from https://www.dailydot.com/irl/twitter-slut-ana-valens/

[17] Sarah Hewitt, T. Tiropanis, and C. Bokhove. 2016. The problem of identifying misogynist language on Twitter (and other online social spaces). In Proceedings of the 8th ACM Conference on Web Science (WebSci '16), Association for Computing Machinery, New York, NY, USA, 333–335. DOI:https://doi.org/10.1145/2908131.2908183

[18] Toxic Twitter - The Reporting Process. Retrieved February 26, 2021 from https://www.amnesty.org/en/latest/research/2018/03/online-violence-againstwomen-chapter-4/

[19] Bailey Poland. 2016. TYPES OF CYBERSEXISM: What Online Harassment Really Looks Like. In Haters: Harassment, Abuse, and Violence Online. University of Nebraska Press, 35–60. DOI:https://doi.org/10.2307/j.ctt1fq9wdp.4

[20] Desmond Upton Patton, Patrick Leonard, Caitlin Elaesser, Robert D. Eschmann, Sadiq Patel, and Shantel Crosby. 2019. What's a Threat on Social Media? How Black and Latino Chicago Young Men Define and Navigate Threats Online. Youth & Society 51, 6 (September 2019), 756–772. DOI:https://doi.org/10.1177/0044118X17720325

[21] Michael Brice-Saddler. He easily found hundreds of death threats against Rep. Ilhan Omar. He wants Twitter to stop them. Washington Post. Retrieved February 26, 2021 from https://www.washingtonpost.com/technology/2019/04/16/heeasily-found-hundreds-death-threats-against-rep-ilhan-omar-he-wantstwitter-stop-them/

[22] Donnie O'Sullivan and Alaa Elassar. Twitter bans posts wishing for Trump death.The Squad wonders where that policy was for them. CNN. Retrieved December 6, 2020 from https://www.cnn.com/2020/10/03/politics/twitter-trump-policy-banthe-squad-politics-trnd/index.html

[23] Eva. 2020. Hands up if someone has ever tweeted wishing for you to die or suffer bodily harm. @evacide. Retrieved April 24, 2021 from https://twitter.com/evacide/status/1312171500087042048

[24] Megan Specia. 2019. Threats and Abuse Prompt Female Lawmakers to Leave U.K. Parliament. The New York Times. Retrieved April 19, 2021 from https://www.nytimes.com/2019/11/01/world/europe/women-parliament-abuse.

[25] Sarah Kerr, Ainara Tiefenthäler, and Nicole Fineman. Video: 'Where's Your
Husband?' What Female Candidates Hear on the Trail. The New York Times. Retrieved March 22, 2021 from https://www.nytimes.com/video/us/politics/100000006027375/women-politics-harassment.html

[26] Sarah Kelly. 2018. Investigating the phonetic and linguistic features used by speakers to communicate an intent to harm. Ph.D.

Dissertation. University of York.

[27] Bruce Fraser. 1998. Threatening revisited. Forensic Linguistics-the International Journal of Speech Language and The Law - FORENSIC LINGUIST 5, (December, 1998), 159–173. DOI:https://doi.org/10.1558/sll.1998.5.2.159

[28] JL Austin. 1962. How to do things with words. Oxford University Press (1962), 174.

[29] Rae Langton. 1993. Speech Acts and Unspeakable Acts. Philosophy & Public Affairs 22, 4 (1993), 293–330.

[30] Kate Storey. 1995. The language of threats. IJSLL 2, 1 (1995), 74–80. DOI:https://doi.org/10.1558/ijsll.v2i1.74

[31] Tammy A. Gales. 2010. Ideologies of Violence: A Corpus and Discourse Analytic Approach to Stance in Threatening Communications. Ph.D Dissertation. University of California, Davis.

[32] Tammy Gales. 2015. Threatening Stances: A corpus analysis of realized vs. nonrealized threats. Language and Law 2, (2015), 25.

[33] Mitchell J. Abrams. 2019. Uncovering The Genre Of Threatening Texts: A Multilayered Corpus Study. Master's thesis. Georgetown University, Washington,DC.

[34] Holger Limberg. 2008. Threats in Conflict Talk: Impoliteness and Manipulation. In Impoliteness in Language: Studies on its Interplay with Power in Theory and Practice, Derek Bousfield, Miriam A. Locher, Ed. 155-179.

[35] Susan Berk-Seligson and Mitchell A. Seligson. 2016. Reported threats: The routinization of violence in Central America. PRAG 26, 4 (December 2016), 583–607. DOI:https://doi.org/10.1075/prag.26.4.03ber

[36] Roger W. Shuy. 1993. Language crimes: The use and abuse of language evidence in the courtroom. Cambridge, MA.

[37] James T. Turner and Michael G. Gelles. 2003. Threat assessment: A risk management approach. Haworth Press, New York, NY.

[38] Violent threats policy. Retrieved February 25, 2021 from https://help.twitter.com/en/rules-and-policies/violent-threats-glorification

[39] Roger W. Shuy. 2005. Creating Language Crimes: How Law Enforcement Uses (and Misuses) Language. Oxford University Press, Cambridge, MA.

[40] What Does Free Speech Mean? United States Courts. Retrieved April 28, 2021 from https://www.uscourts.gov/about-federal-courts/educational-resources/about-educational-outreach/activity-resources/what-does

[41] David L. Hudson Jr. Watts Factors. The First Amendment Encyclopedia: Presented by the John Seigenthaler Chair of Excellence in First Amendment Studies. Retrieved December 26, 2020 from https://www.mtsu.edu/first-amendment/article/1525/watts-factors

[42] Paul T Crane. "True Threats" and the Issue of Intent. Virginia Law Review 92, 55.

[43] Court Role and Structure. United States Courts. Retrieved April 24, 2021 from https://www.uscourts.gov/about-federal-courts/court-role-and-structure

[44] School Shooter. Federal Bureau of Investigation. Retrieved February 25, 2021 from https://www.fbi.gov/file-repository/stats-services-publications-schoolshooter-school-shooter/view

[45] Lawrence M. Solan and Peter M. Tiersma. Speaking of Crime: The Language of Criminal Justice. The University of Chicago Press, Chicago

[46] Mary Margaret Roark. 2015. Elonis v. United States: The Doctrine of True Threats:Protecting Our Ever-Shrinking First Amendment Rights in the New Era of Communication. Pittsburgh Journal of Technology Law & Policy 15, 2 (August 2015), 197–223. DOI:https://doi.org/10.5195/tlp.2015.162

[47] Elonis v. United States, 575 U.S. __ (2015)

[48] Angus Johnston. 2015. Why Today's Elonis Decision is a Victory in the Fight Against Online Harassment. Student Activism. Retrieved November 18, 2020 from https://studentactivism.net/2015/06/01/why-todays-elonis-decision-is-avictory-in-the-fight-against-online-harassment/

[49] John Sivils. 2019. Online Threats: The Dire Need for a Reboot in True-Threats Jurisprudence. SMU L. Rev. F. 72, 1 (November 2019), 51–58. DOI:https://doi.org/10.25172/slrf.72.1.5

[50] Grace Panetta and Samantha Lee. 2019. Twitter is the most popular social media platform for members of Congress — but prominent Democrats tweet more often and have larger followings than Republicans. Business Insider. Retrieved April 28, 2021 from https://www.businessinsider.com/democratic-republican-congresstwitter-followings-political-support-2019-2

[51] Cécile Guerin and Eisha Maharasingam-Shah. 2020. Public Figures, Public Rage: Candidate abuse on social media. ISD: Institute of Strategic Dialogue. Retrieved April 28, 2021 from https://www.isdglobal.org/isd-publications/public-figurespublic-rage-candidate-abuse-on-social-media/

[52] MEET YOUR MAKER (phrase) American English definition and synonyms | Macmillan Dictionary. Retrieved February 26, 2021 from https://www.macmillandictionary.com/us/dictionary/american/meet-your-maker

[53] Examples of Hyperbole: What It Is and How to Use It. Retrieved February 26, 2021 from https://examples.yourdictionary.com/examples-of-hyperboles.html

[54] Ryan Lizza, Daniel Lippman, and Meridith McGraw. AOC wants to cancel those who worked for Trump. Good luck with that, they say. POLITICO. Retrieved February 26, 2021 from https://www.politico.com/news/2020/11/09/aoc-cancelworked-for-trump-435293

[55] Trump Accountability Project: Meet new group seeking to blacklist staff who worked for Trump administration - World News , Firstpost. Retrieved February 26, 2021 from https://www.firstpost.com/world/trump-accountabilityproject-meet-new-group-seeking-to-blacklist-staff-who-worked-for-trumpadministration-9007641.html

[56] Definition of HUNT. Retrieved February 26, 2021 from https://www.merriamwebster.com/dictionary/hunt

[57] Davey Alba, Kate Conger, and Raymond Zhong. 2020. Twitter Adds Warnings to Trump and White House Tweets, Fueling Tensions. The New York Times. Retrieved April 28, 2021 from https://www.nytimes.com/2020/05/29/technology/trumptwitter-minneapolis-george-floyd.html

[58] Who Is AOC: Alexandria Ocasio-Cortez on Her Rise to Political Power | Vanity Fair. Retrieved February 26, 2021 from https://www.vanityfair.com/news/2020/10/becoming-aoc-cover-story-2020

[59] Breaking the News: Censorship, Suppression, and the 2020 Election | United States Senate Committee on the Judiciary. Retrieved February 26, 2021 from https://www.judiciary.senate.gov/meetings/breaking-the-news-censorshipsuppression-and-the-2020-election

[60] Nassim JafariNaimi, Lisa Nathan, and Ian Hargraves. 2015. Values as Hypotheses: Design, Inquiry, and the Service of Values. Design Issues 31, (October 2015), 91–104. DOI:https://doi.org/10.1162/DESI_a_00354

Pooja Casula

Pooja Casula is a Ph.D. student in the Digital Media program working with Dr. Nassim Parvin. Her research interest lies at the intersection of social media, tech policy, and politics, specifically in understanding how these influence democracy and political participation. Her current project explores gendered abuse and disinformation campaigns on social media targeting women in politics.

Dr. Aditya Anupam

Aditya Anupam is a Postdoctoral Researcher at the School of Literature, Media, and Communication at Georgia Tech. He works on Ethics, Technology, and Education as part of the Design and Social Justice Studio led by Dr. Nassim Parvin. Aditya received his Ph.D. in Digital Media in December 2021 from the same department. His research is situated at the confluence of science, media, and learning. Anchored in feminist, STS, and pragmatist scholarship, he explores digital media––particularly games, simulations, and interactive visualizations––as environments to foster the learning of science as a situated practice

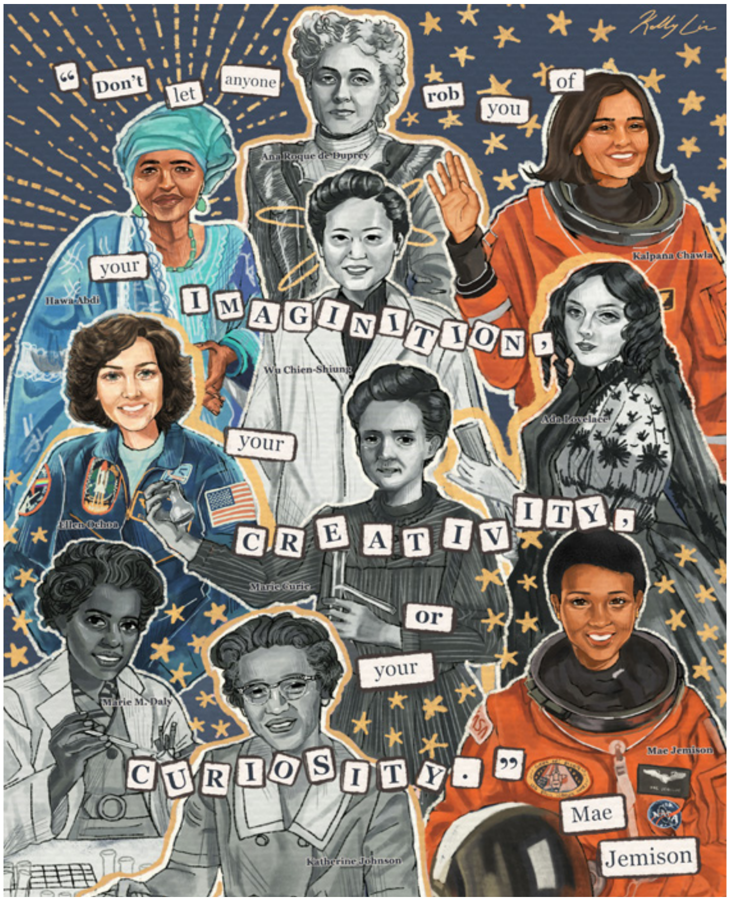
Dr. Nassim Parvin

Dr. Nassim Parvin is an Associate Professor at the School of Literature, Media, and Communication at Georgia Tech, where she also serves as an associate director to the Digital Integrative Liberal Arts Center (DILAC). Dr. Parvin's research explores the ethical and political dimensions of design and technology, especially as related to questions of democracy and social justice. Dr. Parvin's interdisciplinary research integrates theoretically-driven humanistic scholarship and design-based inquiry. Her scholarship has appeared in premier publication venues in design studies, science and technology studies, and human-computer interaction. Her designs have been deployed at non-profit organizations such as the Mayo Clinic and exhibited in venues such as the Smithsonian Museum, receiving multiple awards and recognitions. She is one of the lead editors of Catalyst: Feminism, Theory, Technoscience, an award-winning open-access journal in the expanding interdisciplinary field of STS and serves on the editorial board of Design Issues. Dr. Parvin's teaching has also received multiple recognitions inclusive of the campus-wide 2017 GATECH CETL/BP Junior Faculty Teaching Excellence Award.

Kelly Lin

Kelly is a first-year Computational Media student with a concentration in People and Interaction Design. They are a self-taught artist and designer who is passionate about inclusion and expression. They strongly believe that both art and technology act as bridges between our personal world and the people around us; they are crucial elements of human connection and communication.

The Named Pioneers



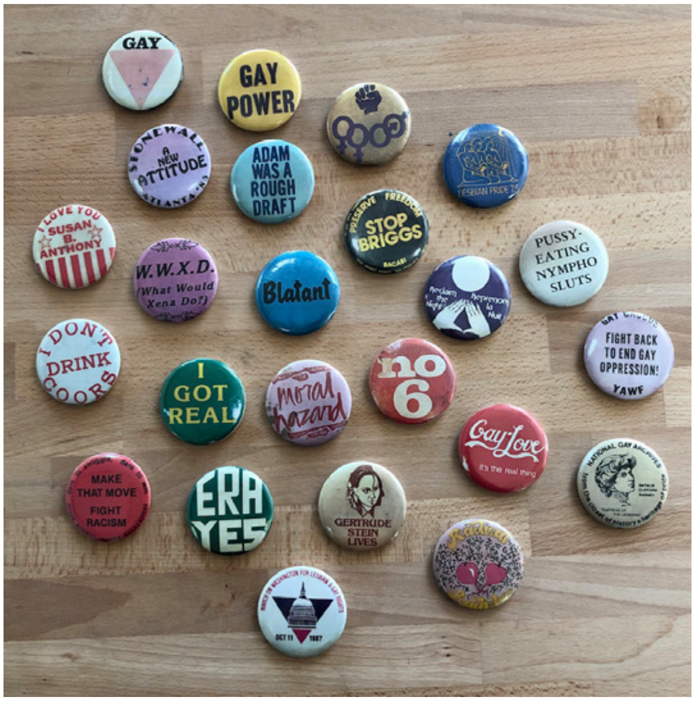
The Named Pioneers: Katherine Johnson, a NASA mathematician during the time of the first crewed spaceflights, has only recently been recognized by the movie, "Hidden Figures." Johnson was nameless in her profession dominated by white men, despite how crucial her calculations were for the advancement of technology. This is the unfortunate reality for marginalized communities, whose countless contributions are constantly overlooked. The celebration and representation of these communities is extremely important for any aspect of our society, and it's necessary for both innovation and human potential. Hawa Abdi: a trailblazing physician and human rights activist from Somalia. Ana Roqué de Duprey: a prolific academic, educator, and suffragist from Puerto Rico. Kalpana Chawla: is an esteemed aerospace engineer and the first Indian woman in space. Ellen Ochoa: a researcher, optical systems engineer, and the first Latina woman in space. Wu Chien-Shiung: a researcher, professor, and pivotal figure in physics, particularly atomic science. Ada Lovelace: a visionary mathematician, regarded as the first computer programmer. Marie M. Daly: the first black woman to receive a Ph.D. in chemistry in the United States, and an activist for minority academic rights. Marie Curie: a physicist known for her work in radioactivity, and the first and only woman to receive the Nobel Prize twice. Mae Jemison: a physician and chemical engineer, was the first black woman to travel into space. This is a tribute to these exceptional women, the few of many, for, not only their dedication and determination as pioneers in their passions but also their resilience as women.

Alexandra Teixeira Riggs

Alexandra Teixeira Riggs (she/they) is a PhD student in Digital Media at Georgia Tech, concentrating in queer media studies, critical making, and design justice. Their work focuses on using interactive storytelling methods to explore both past and present notions of queer community, identity, and belonging. They combine both tangible and screen-based interfaces with design research to challenge dominant technologies and offer alternative relationalities in both on and offline space. They are currently working with Dr. Anne Sullivan in the Storycraft Lab, and Dr. Noura Howell. For their current research, they are also working with archivist Morna Gerrard in the Gender and Sexuality Collection at Georgia State University.

Lorraine, Maria Helena, and Charlene

Lorraine, Maria Helena, and Charlene is an artifact portrait of three queer activists who influenced and produced Atlanta's patchwork of LGBTQ+ organizations from the mid-1970s to the late 1990s (and ongoing). The composite image of each individual's collection of buttons and pins is part of a larger research project entitled Button Portraits: Embodying Queer History with Interactive Wearable Artifacts, which seeks to represent queer history through tangible interactive narrative. The project asks, "How do we reflect on our history, and what bearing does it have on our current conceptions of queer community, identity, and belonging?" In Button Portraits, I curate a selection of button artifacts, self-reflexively acknowledging my own involvement in shaping a story and representing history. The full narrative piece uses these buttons as tangible interactive objects that reveal fragments of oral histories, sourced from interviews with two activists, Lorraine Fontana and Maria Helena Dolan. In the piece, placing a physical button on your chest will complete a capacitive circuit to play a fragment of oral history, and each new button held will unlock a new narrative piece. The gesture of holding a button to your chest, pinning it, and intimately listening to an interview with the button's original owner serves to not only implicate participants in history, but also to reframe our relationships to stories through attention to the bodily experience. By designing this interaction, I explore how we might question and reflect on the unknowable and partial nature of history, on our own identities in relation to one another, and on our communities both past and present.


GAY
GAY POWER
STONEWALL A NEW ATTITUDE ATLANTA
ADAM WAS A ROUGH DRAFT
PRESERVE FREEDOM STOP BRIGGS
LESBIAN PRIDE
I LOVE YOU SUSAN B. ANTHONY
W.W.X.D. (What Would Xena Do?)
Blatant
Reclaim the Night
PUSSY-EATING NYMPHO SLUTS
I DON'T DRINK COORS
I GOT REAL
moral hazard
no 6
FIGHT BACK TO END GAY OPPRESSION! YAWF
Gay Love it's the real thing
MAKE THAT MOVE FIGHT RACISM
ERA YES
GERTRUDE STEIN LIVES
OCT 11 1987
NATIONAL GAY ARCHIVES

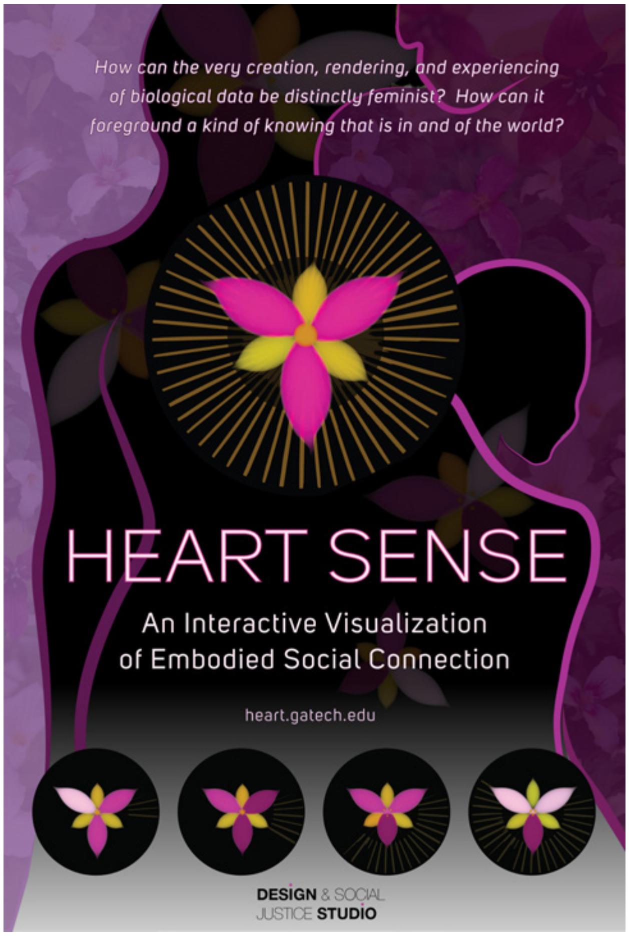
Sylvia Janicki

Sylvia Janicki is a first-year PhD student in Digital Media leading efforts in the design and development of the current iteration of Heart Sense. Sylvia has a background in landscape architecture. Her research centers on issues of access and justice in urban environments and explores the intersections of built, digital, and bodily spaces. She is currently working with Dr. Nassim Parvin in the Design and Social Justice Studio to examine embodied experiences of sensing and data production with implications for designing affective technologies in smart cities.

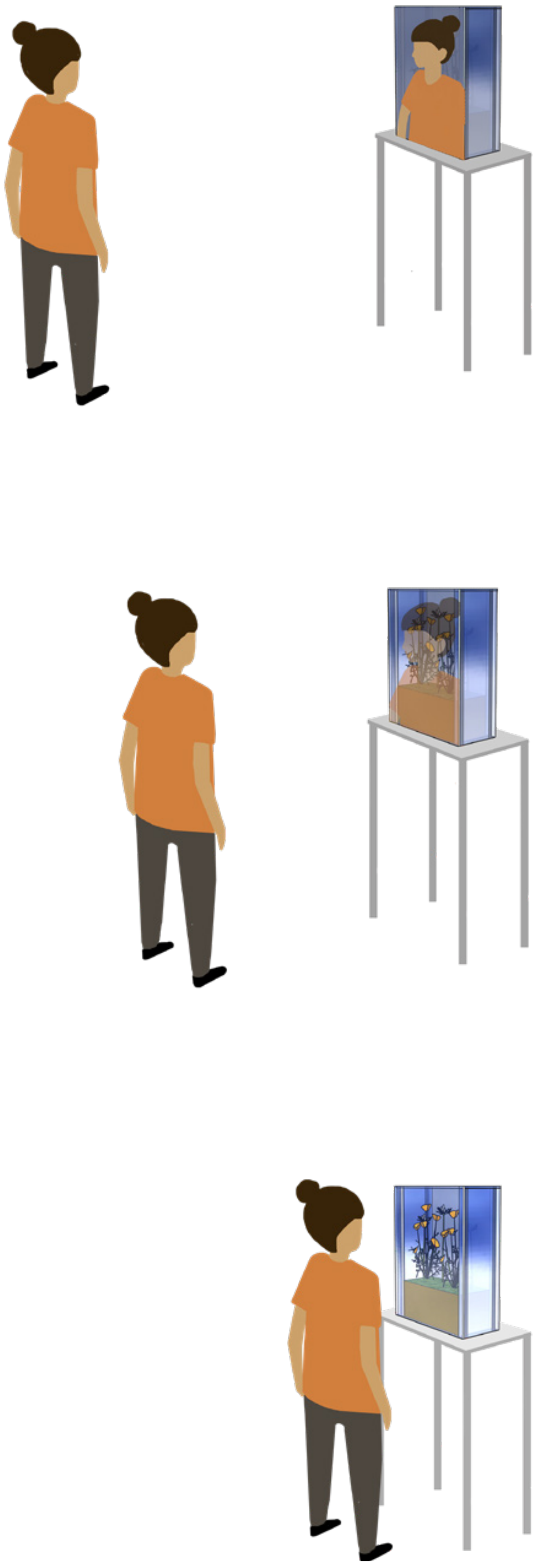
Heart Sense

Heart Sense is led by Dr. Nassim Parvin (Digital Media, Georgia Tech), Anne Pollock (Global Health and Social Medicine, Kings College London), Lewis Wheaton (Biological Sciences, Georgia Tech). Student collaborators for Heart Sense include Aditya Anupam, Shubhangi Gupta, and Mohsin Yousufi

Heart Sense is a series of art installations that use representation, tracking, and visualizations of physiological data to investigate and reflect upon the body in ways that depart from the quantitative self and spur curiosity about scientific measurements.
How can the very creation, rendering, and experiencing of biological data be distinctly feminist? How can it contribute to a more nuanced understanding of our bodies? How can it foreground a kind of knowing that is in and of the world? How can it break down binaries that have been the subject of criticism in the sciences such as objectivity and subjectivity; self and other; individual and social?
In this iteration, the installation engages with social dimensions of embodiment through the mediation of the physical environment. Three participants are invited to sit around a table and are given headphones to listen to music. A floral visualization representing both individual and collective heart rates of the participants will be projected onto the table, the size and the colors of each petal shifting with changes in each participant's body. The visualization showcases how our bodies come into relation with each other and are in and of the environment, as they respond to our surrounding conditions even when we are not aware of it. The floral form takes inspiration from the trillium, a spring ephemeral whose individual flowering bodies are connected by a system of underground rhizome roots.


How can the very creation, rendering, and experiencing of biological data be distinctly feminist? How can it foreground a kind of knowing that is in and of the world?
HEART SENSE
An Interactive Visualization of Embodied Social Connection
heart.gatech.edu
DESIGN & SOCIAL JUSTICE STUDIO

A Memorial to Earth



How can living materials be combined with digital technologies to reframe the body-environment dichotomy, challenge bounded individualism, and elevate care for more than human communities? This project draws from scholarship in ecofeminism and feminist post-humanities that contextualizes ecological destruction in the modern landscape to reframe human and non-human worlds as entangled, people and environment as co-constitutive, and to foreground care for and agency of more-than-human communities.

This memorial consists of living wildflowers growing inside a plexiglass box. The box is lined with a two-way mirror on the front, a regular mirror on the back, and LED strip lights along the sides. An ultrasonic sensor senses distance of a body from the memorial, triggering LED lights to brighten when the body approaches, and dim when the body recedes. At a distance, with LED lights dimmed or off, one would see only their own reflection. As one approaches, LED lights brighten, creating a co-constitutive view of the person becoming one with the plants. When the viewer is directly in front of the memorial, their reflection disappears, leaving only a view of the wildflowers and the soil within, illuminated under an infinity mirror effect, constructing an illusion of an infinite meadow, signaling the limits of the Earth.

Finally, participants are encouraged to open the lid to feel and water the plants. This memorial serves as an invitation to attend to and care for the more-than-human world, recognizing our accountability to non-human species. The installation places the human, plants, and earth in an entangled network, connected by digital technologies. The memorial may seem closed and bounded, but is, in fact, an open system, necessitating porosity and exchange with the outside and care from humans in order to thrive.

Rethinking Safe Mobility: The Case of Safetipin in India

Shubhangi Gupta Sylvia Janicki Pooja Casula Dr. Nassim Parvin

INTRODUCTION

Around the world, women, along with other marginalized groups, experience violence and abuse in their homes and on the streets. India is no exception. To promote safety in cities, the government and technology companies have created tools such as tracking apps, reporting apps, and alarm systems. However, these initiatives and technologies have been criticized for making gendered assumptions about safety. In our research, we draw upon the work of feminist scholars, urban planners, and sociologists, especially qualitative and critical studies on safety in cities to identify four key themes surrounding criticisms of safety technologies namely: (1) they project the fear of assault onto the urban environment; (2) they put the responsibility of being safe on women; (3) they enable surveillance and control; and (4) they disregard intersectionality (they fail to account for the influence of age, gender, ability, class, caste, race, and religion on safety). Together, these factors exacerbate the challenges of safe mobility.

In contrast to the criticisms received by safety technologies, another such tool named Safetipin has received broad appreciation. The app has won several awards and has been included in a variety of ‘best safety apps’ lists. Given the appreciation for Safetipin, it is essential to examine whether and how the app withstands the criticisms that have been raised against other safety tools. Is it successfully addressing those criticisms? Or, does it bring about the same issues under a new guise?

EXAMINING SAFETIPIN’S STRENGTHS AND SHORTCOMINGS

Safetipin is a mobile application created by a group of women’s rights advocates in India following the aftermath of the horrific Nirbhaya rape case in 2012. The app allows users to share data about a location’s infrastructure by rating the following nine factors: amount of lighting, openness of space, visibility to others, number of people around, presence of security, condition of walk paths, availability of public transit, presence of women and children, and feeling of safety. This crowdsourced data is used to calculate safety scores for various locations and routes. A user can then review these safety scores to make safe mobility decisions. The Safetipin app also allows the user to share their location with others. According to the app’s website, the crowdsourced data is shared with the local government to improve city infrastructure such as lighting and walk paths. Through these practices, Safetipin seeks to help its users safely navigate the urban environment.

Below we discuss if or how Safetipin addresses or fails to address the four key criticisms against other safety technologies summarized above.

Projecting Fear Onto the Urban Environment

Women commonly report fearing strangers, especially at night and in public spaces. Safetipin aims to assuage these fears by informing women of ‘safe’ and ‘unsafe’ routes. However, in doing so, Safetipin further advances the fear of ‘unsafe’ spaces, limiting women's mobility to 'safe' spaces. 'Unsafe' spaces, then, draw fewer women, further lowering the areas' safety scores. Even as women walk on paths marked 'safe', the less-than-perfect rating continues to frame assault as an ever-present possibility.

Another way Safetipin aims to mitigate fear is by partnering with the Indian government to improve public infrastructure such as street lights. However, fear cannot simply be 'designed out' through such improvements. Instead, it needs to be addressed alongside social, legal, and economic infrastructures. Moreover, public officials in charge of making pragmatic use of the safety data may not take a progressive stance on women's safety. Instead, they may misuse safety data to advance sexist practices such as banning 'provocative' clothing.

Placing Responsibility on Women

The idea of 'respectability' dictates that 'good' women act decently and have a purpose to be outside of their homes. 'Respectability' is often operationalized through cultural and political narratives as a way to prevent sexual assault. On the surface, it may seem that Safetipin shifts such narratives by bringing awareness to issues of safety and empowering women to make informed mobility decisions. However, the same routes that are labeled as 'safe' have the potential to become yet another ideal for women to adhere to. 'Respectable' women may be expected to limit their movement to 'safe' spaces, and their presence in 'unsafe' areas may be considered a transgression. With the use of Safetipin, adherence to safety scores then becomes the new respectable behavior that will keep women safe. This approach puts the burden of safety on women and ultimately does little to address the root causes of violence.

Enabling Surveillance and Cyber Control

Women commonly express feeling constrained by their family members' constant watch over them. At the same time, however, they express the need to be in contact with someone they trust when they are in public. Due to the fear of assault in public spaces, they accept the compromise of freedom for alleged safety. Safetipin internalizes this compromise by allowing one to share their location with others. Further, such a feature positions women as weak and dependent and requires them to surrender control of their bodies. This feature can also be abused to stalk or control women's movement and exacerbate domestic violence. It may appear that using this feature is a choice that women make, as suggested by Safetipin in our interviews. However, the ingrained fear of assault, the respectability narratives, and the social and familial structures strongly influence this 'choice'.

Disregarding the Intersectional Nature of Safe Mobility

Both experiences and perceptions of safety and mobility are intersectional. They are entangled with systems of oppression such as gender, ability, age, race, caste, and religion. Safetipin attempts to be inclusive of diverse safety experiences by providing alternate modes of technological access. For example, they set up safety centers to address the needs of people who may not have access to a mobile phone. This diverse data is used to aggregate safety risks in the form of a universal, objective 'safety score'.

However, in their attempt to calculate an objective measure of safety, Safetipin reduces the diversity and specificity of safety experiences. They disregard the importance of the data setting in which each data point is created: Who created the data? Whose safety was calculated? In whose company? Under what circumstances?

Seemingly 'objective' depictions of safety could still be based on prejudiced conceptions of who is dangerous. As such, the app can act as a means to reify unjust assumptions that neighborhoods of lower socioeconomic status and minority religions are unsafe. Such unfounded suggestions risk further reinforcing fear of and stigma against marginalized groups while disregarding the presence of violence at home by known perpetrators.

ADVANCING SAFETY, CHALLENGING PATRIARCHY

The inadequacies and failures of safety tools exemplify how difficult it is to break free from dominant patriarchal norms that seep into the design of emerging technologies and reinforce long-standing injustices. While it may be said that Safetipin has initiated conversations around an important issue, its impact and efficacy remain unclear.

"But we cannot risk women's lives just to challenge the patriarchy!" is a common reprise. We wonder, however, if the space of possibility is limited to the binary choice stated above. Are our only options either to risk lives to challenge patriarchy or to save lives by advancing patriarchal norms? Alternatively, how can we, as ICTD designers, do both: design for women's safety while challenging patriarchy? To advance safety in a meaningful manner, we need not only be aware of the discriminatory nature of these approaches but also actively protest them. Nuanced and critical examination of safety technologies and their narratives marks a vital first step in this process.

The full paper discussing this work will be released as part of the 12th International Conference on Information & Communication Technologies and Development (ICTD 2022).

Shubhangi Gupta

Shubanghi Gupta is a Ph.D. student in Digital Media at the Georgia Institute of Technology, advised by Dr. Nassim Parvin. In her research, she draws upon feminist and STS scholarship integrated with qualitative research methods to explore questions of safety and social justice as they relate to the design of emerging technologies. Contact: shubhangi@gatech.edu

Sylvia Janicki

Sylvia Janicki is a first-year PhD student in Digital Media leading efforts in the design and development of the current iteration of Heart Sense. Sylvia has a background in landscape architecture. Her research centers on issues of access and justice in urban environments and explores the intersections of built, digital, and bodily spaces. She is currently working with Dr. Nassim Parvin in the Design and Social Justice Studio to examine embodied experiences of sensing and data production with implications for designing affective technologies in smart cities.

Pooja Casula

Pooja Casula is a Ph.D. student in the Digital Media program working with Dr. Nassim Parvin. Her research interest lies at the intersection of social media, tech policy, and politics, specifically in understanding how these influence democracy and political participation. Her current project explores gendered abuse and disinformation campaigns on social media targeting women in politics.

Dr. Nassim Parvin

Dr. Nassim Parvin is an Associate Professor at the School of Literature, Media, and Communication at Georgia Tech, where she also serves as an associate director to the Digital Integrative Liberal Arts Center (DILAC). Dr. Parvin's research explores the ethical and political dimensions of design and technology, especially as related to questions of democracy and social justice. Dr. Parvin's interdisciplinary research integrates theoretically-driven humanistic scholarship and design-based inquiry. Her scholarship has appeared in premier publication venues in design studies, science and technology studies, and human-computer interaction. Her designs have been deployed at non-profit organizations such as the Mayo Clinic and exhibited in venues such as the Smithsonian Museum, receiving multiple awards and recognitions. She is one of the lead editors of Catalyst: Feminism, Theory, Technoscience, an award-winning open-access journal in the expanding interdisciplinary field of STS and serves on the editorial board of Design Issues. Dr. Parvin's teaching has also received multiple recognitions inclusive of the campus-wide 2017 GATECH CETL/BP Junior Faculty Teaching Excellence Award.

There can be no sustainable future without gender equality