Motivation. Hate is a contagion that can spread by contact. I have observed evidence of this in my own work [1]. Xenophobia and fear spread across social media like an ugly shadow behind the COVID-19 outbreak, inciting both verbal and physical violence towards individuals of Asian descent. The proper response to these and other online abusive behaviors, I believe, bears a striking resemblance to a collective pandemic response: that is, the most effective measures are *preventative*. If we identify asymptomatic "carriers" *before* they have published their first harmful post, then we can encourage these individuals to "quarantine" any messages they had drafted and reflect on their risk to the community. Furthermore, we can warn community members before they engage with any potentially toxic or abusive messages.

Previous efforts to detect, flag, and then *later* remove toxic language all miss this critical window for *averting harm* and *preventing spreading.* Furthermore, existing classifiers have a limited capacity to detect only a narrow range of explicitly toxic language containing profane, offensive, or racially-charged language. These classifiers overlook the more *subtly harmful language* of dog-whistles, implicit hate speech, and microaggressions, as well as *harmful behaviors* like doxxing, trolling, and cyberbullying, which can be repetitive and are largely determined by norms and social context, which have not been represented in the training data [2]. I will address each of these shortcomings in turn. In the next three years of my PhD, I will:

1. **Design** a *comprehensive taxonomy* of online abuse in collaboration with social scientists and other domain experts
2. **Develop** *computational sociolinguistic models* of online abuse
3. **Collect** a more nuanced and challenging set of *benchmark datasets* with rich user meta-data to represent the sociolinguistic variables relevant to my social models
4. **Build** a *proactive intervention tool* to predict and prevent online abuse before it unfolds

Designing a comprehensive taxonomy. Previously, I designed a novel cyberbullying detection framework [2] based on the social sciences literature on bullying. My framework included a codebook for inferring not only aggressive language, but also harmful intent, repeated action, public visibility and the imbalance of power that distinguish cyberbullying from other forms of abuse. Now, in collaboration with domain experts, I will extend this work with a *comprehensive taxonomy* of online toxicity, outlining the principal axes of variation that may render a catch-all "abuse" classifier infeasible. My taxonomy will inform better computational models, tailored to specific sociolinguistic parameters that define each toxicity category.

One critical distinction that I will make is between *denotation* and *connotation*. Explicit hate speech is a *denotative* expression of derogatory views that one might reasonably expect to detect using a text-classification model. Dog-whistles and microaggressions, on the other hand, are *connotative*. They implicitly reflect the speaker's social biases and ideology, which may not be signaled in the text itself. Furthermore, hate speech is defined *semantically*, while other forms of abuse like trolling and cyberbullying are defined *pragmatically* by anti-cooperative behaviors and norm violations, both Grician and community-specific. These and other similar insights will guide the modeling assumptions of future work, since a model that is trained on isolated message text alone cannot infer microaggressive connotations or the pragmatic dimensions of hate.

Developing theoretically-grounded models. Once I have established a comprehensive abuse taxonomy, it will directly guide my own modeling decisions. For abuse categories that are determined by *discourse structure*, I will experiment with a recursive neural network architecture over a dependency-like rhetorical structure (RST) tree. Because RST structure captures contrast, elaboration, concession, and antithesis between units of discourse, I can expect my model to distinguish messages that elaborate on shared frustration ("yeah, f*** climate change"), and messages that express retaliation ("f*** you"). To capture norms and power dynamics, I will engineer new features from users' social networks and conversation histories, and I will incorporate structured knowledge from sentic knowledge bases to infer the implicit connotations that messages carry. In later projects, I will work to reconcile the complex dialogic interactions between multimodal discourse elements that give rise to outcomes like irony and humor. In summary, I will contribute a rich set of computational sociolinguistic models across different forms of abuse, with features ranging over the relevant social, psychological, and linguistic variables that shape and define civility in online interactions.

Collecting nuanced benchmark datasets. The field still lacks a set of reliable ground-truth datasets on which to benchmark progress. Moreover, previous data annotation efforts have been largely atheoretic, with untrained workers deciding labels for anonymized and isolated text. I will create new benchmark datasets that are grounded in the social science theory of my proposed taxonomy. In particular, I will train annotators to account for *dialect variation* as well as for *social pragmatic* signals in the surrounding message thread. This effort is crucial since untrained workers have been prone to mislabel isolated messages [2] and in doing so, they have amplified racial biases, interpreting African American English (AAE) as toxic [3].

Building a proactive system. Detecting and removing toxic comments is *reactive* rather than *proactive*. My goal is to predict abusive behaviors, and help steer declining conversations before they have derailed. Recent work predicts conversational failure using logistic regression over a collection of hand-engineered *politeness* features [4]. However, even on a very restricted domain (collaborative Wiki edit discussion forums) the classification accuracy is, at best, 65% on a balanced test set. I will extend this work with *user-specific* and *relationship-specific* measures of norms, connotations, and discourse structure as outlined above. Furthermore, I will improve these results by considering, for the first time, *ideological difference* as a *causal* source of disagreement on the web. I have already begun to investigate the way different media outlets frame police shooting incidents, highlighting select discourse elements to promote particular interpretations in their readers. I believe that differences in framing will also reliably signal ideological differences in discourse, and help predict the induced conflict. Once I can predict such conflict, I will build an intervention system to prevent online abuse before it can spread.

Intellectual merit. Online toxicity is a complex and multi-faceted problem that has evaded clear solutions. My research agenda will address three major barriers to progress: (1) I will establish an *underlying theory*, uniting disparate research efforts and informing new computational models of abuse; (2) I will provide *novel benchmark datasets* to measure progress in the field; and (3) I will build *theoretically grounded* and *predictive* models on the rich social, psychological, and linguistic variables that shape the dynamics of civility in online interactions. From a methods standpoint, my work will elucidate the *social dynamics* of language for the field of NLP, which has historically focused only on language *structure* and *literal meaning.*

Broader impact. Hate speech is the catalyst for both *psychological* violence on the web and *physical* violence in the real world, reinforcing systems of oppression and disproportionately impacting minority and disadvantaged communities [3]. My work will help make the internet a more civil and a more welcoming place for everyone, and in doing so, it will also help positively impact and shape the physical spaces of our global and interpersonal lives.

References [1] **Ziems**, et al. (2020). *Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis.* arXiv preprint. [2] **Ziems** et al. (2020). *Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification*. In ICWSM. [3] Sap et al. (2019). *The risk of racial bias in hate speech detection.* In ACL. [4] J. Zhang, et al. (2018). *Conversations gone awry: Detecting early signs of conversational failure*. In ACL.