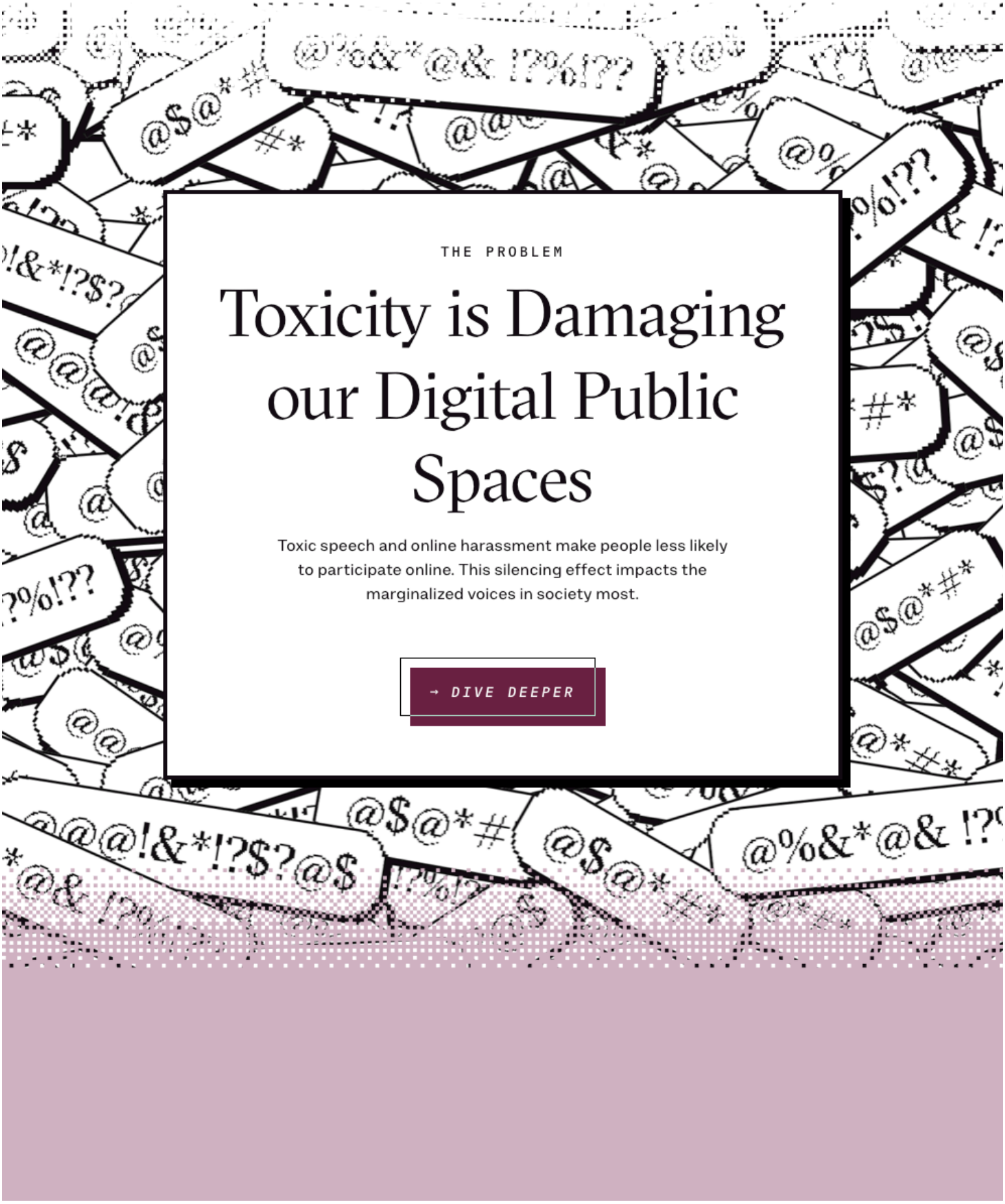
With millions of people spending more time online than ever before, our digital spaces have become a critical forum to engage with one another, share our opinions and participate in the virtual public square. Not so long ago, the promise of the internet was this vibrant space to exchange ideas—one that could facilitate learning and connectedness on a global scale. Much of that promise has been achieved, but this open exchange can be a mixed story. As discussion and expression give way to bullying and berating, toxicity has become a pervasive force in our conversations online.

More than 4 in 10 Americans have experienced online harassment. Globally, minorities—especially people of color and women—are the most likely to experience some form of violence online.

Source: State of Online Harassment

One trip to many websites' comments section and you're likely to see it first-hand. Toxicity may range from overt forms like abusive language and bullying to subtler methods. This behavior infiltrates virtually every corner of the internet, but can be especially pervasive in gaming, news, blogging, and social media.

Safeguarding healthy discourse online ensures everyone has the ability to participate online. Toxicity online imperils the ability of individuals to experience the full promise of the internet.

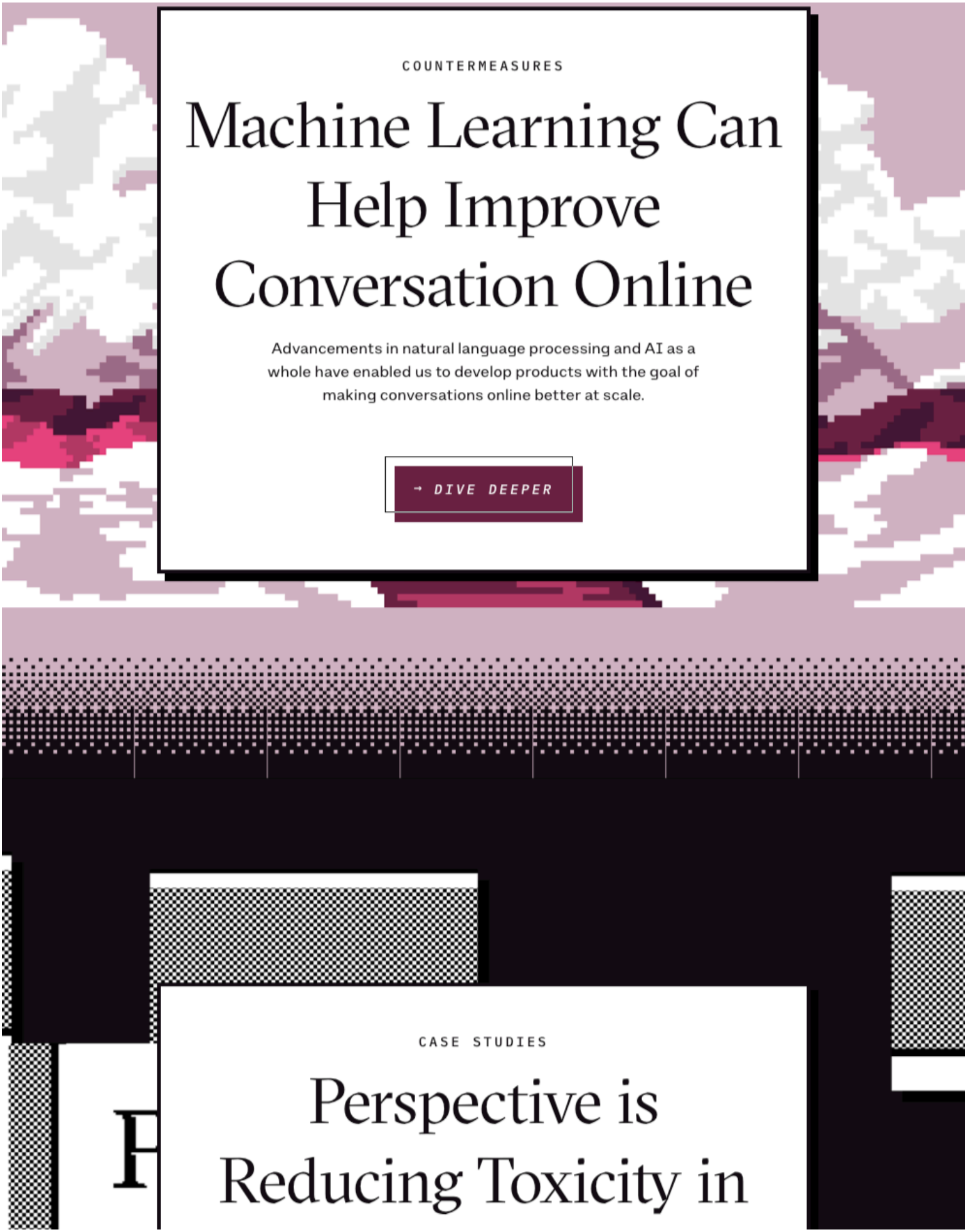


THE PROBLEM

Toxicity is Damaging our Digital Public Spaces

Toxic speech and online harassment make people less likely to participate online. This silencing effect impacts the marginalized voices in society most.

→ DIVE DEEPER


COUNTERMEASURES
Machine Learning Can Help Improve Conversation Online
Advancements in natural language processing and AI as a whole have enabled us to develop products with the goal of making conversations online better at scale.
→ DIVE DEEPER
CASE STUDIES
Perspective is Reducing Toxicity in

the Real World

Jigsaw's API, Perspective, is helping platforms of all sizes, from the New York Times to FACEIT, support healthy, engaging and productive conversations.

→ DIVE DEEPER

Why we use the term "toxicity"

Throughout this report we use the term "toxicity," defined as rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion. We specifically use the term toxicity, as opposed to words like "abuse" or "hate speech" for both practical and scientific reasons.

First, toxicity refers to a broad category of language that is subject to individual interpretation. Terms like "hate speech" or "abuse" frequently refer to specific categories of language that violate certain terms of service or laws. Different platforms or countries have different rules that govern these specific categories of speech. Our research and technology are focused on a broader category of language online, so that they can be useful to a wide range of online forums which may each have their own policies.

There is a scientific basis for using the term toxicity as well. Part of the machine learning development

process behind Perspective requires “training” the artificial intelligence to understand what constitutes harmful language by processing data from millions of public online comments that have been “tagged” by human annotators as having certain features (e.g. racism, misogyny, threats, etc.). Our research suggested that toxicity is an easy concept for annotators to understand, meaning we can gather opinions from a diverse range of people, allowing us to capture the inherent subjectivity of the concept. In other words, it was easier for more people to agree on what constituted “toxic” speech—comments likely to make someone leave a conversation—than it was for people to agree on other terms to describe problematic comments.

We understand that every publisher or platform—and every individual, for that matter—has a slightly different interpretation of what comments are acceptable for them. That’s precisely why we’ve used such an inclusive term for our own research, and why we’ve created Perspective to be flexible and customizable, so each publisher and platform can decide for themselves how best to use this technology.