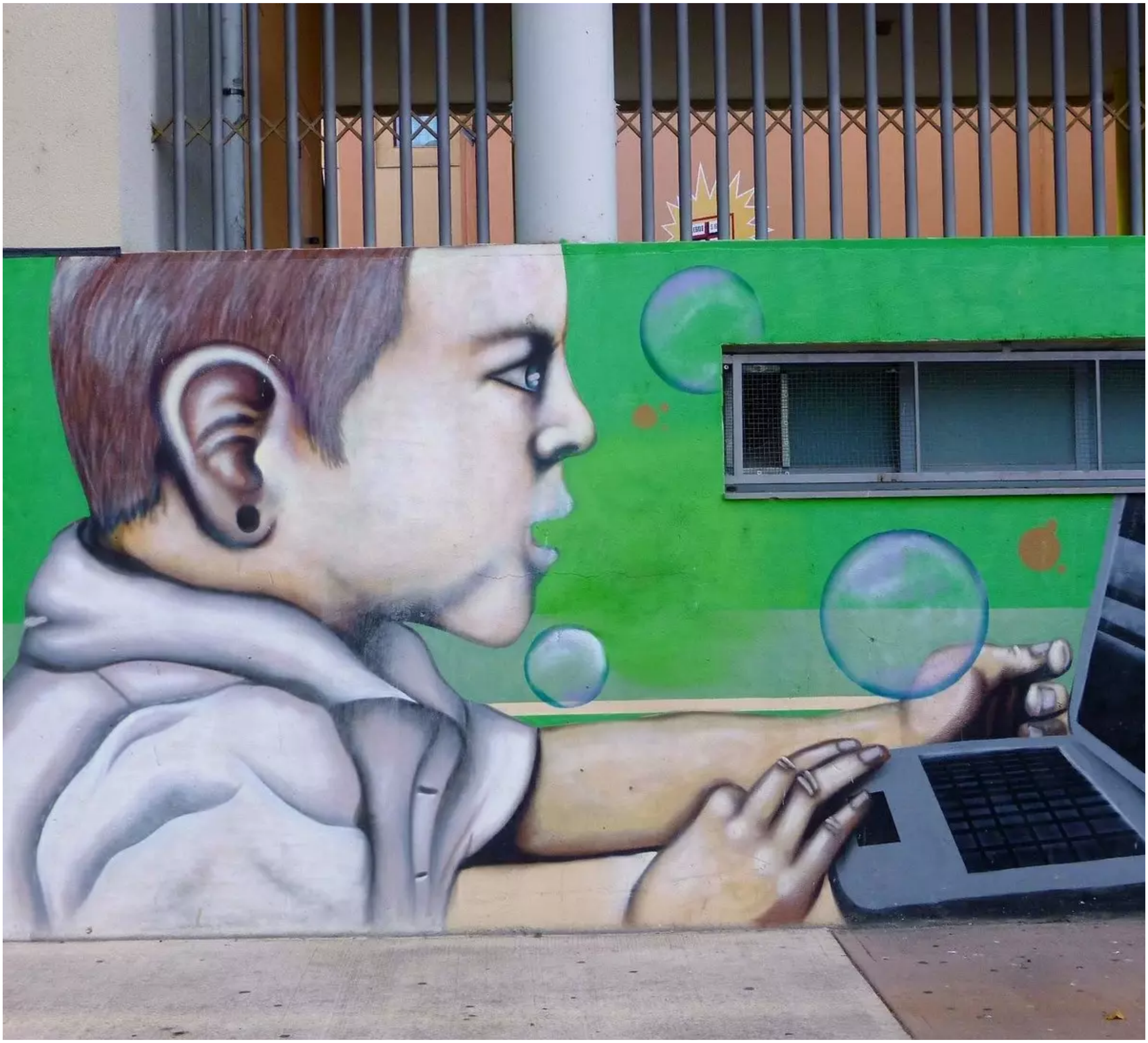
WORLD ECONOMIC FORUM

Join us

CYBERSECURITY

The solution to online abuse? AI plus human intelligence

OPINION Aug 10, 2022



With AI and human intelligence, scaled detection of online abuse can reach near-perfect precision.
Image: Wikimedia/Zarateman

Inbal Goldberger

VP of Trust and Safety, ActiveFence

○ **Listen to the article** **8 min listen**

Readers: Please be aware that this article has been shared on websites that routinely misrepresent content and spread misinformation. We ask you to note the following:

1) The content of this article is the opinion of the author, not the World Economic Forum.
2) Please read the piece for yourself. The Forum is committed to publishing a wide array of voices and misrepresenting content only diminishes open conversations.

With 63% of the world's population online, the internet is a mirror of society: it speaks all languages, contains every opinion and hosts a wide range of (sometimes unsavoury) individuals.

As the internet has evolved, so has the dark world of online harms. Trust and safety teams (the teams typically found within online platforms responsible for removing abusive content and enforcing platform policies) are challenged by an ever-growing list of abuses, such as child abuse, extremism, disinformation, hate speech and fraud; and increasingly advanced actors misusing platforms in unique ways.

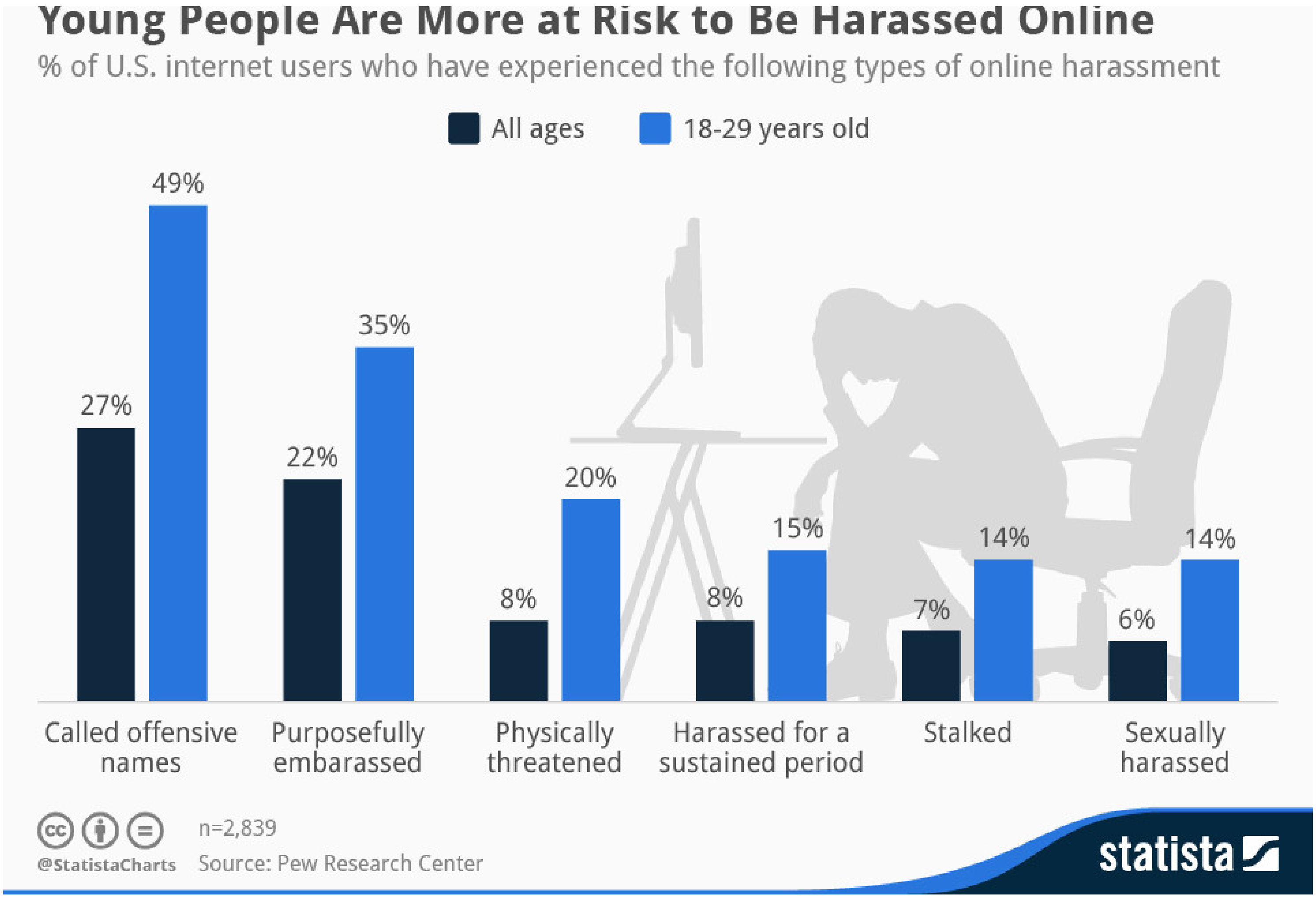
The solution, however, is not as simple as hiring another roomful of content moderators or building yet another block list. Without a profound familiarity with different types of abuse, an understanding of hate group verbiage, fluency in terrorist languages and nuanced comprehension of disinformation campaigns, trust and safety teams can only scratch the surface.

A more sophisticated approach is required. By uniquely combining the power of innovative technology, off-platform intelligence collection and the prowess of subject-matter experts who understand how threat actors operate, scaled detection of online abuse can reach near-perfect precision.

Online abuses are becoming more complex

Since the introduction of the internet, wars have been fought, recessions have come and gone and new viruses have wreaked havoc. While the internet played a vital role in how these events were perceived, other changes – like the radicalization of extreme opinions, the spread of misinformation and the wide reach of child sexual abuse material (CSAM) – have been enabled by it.

Online platforms' attempts to stop these abuses have led to a Roadrunner meets Wile E. Coyote-like situation, where threat actors use increasingly sophisticated tactics to avoid evolving detection mechanisms. This has resulted in the development of new slang, like child predators referring to "cheese pizza" and other terms involving the letters c and p instead of "child pornography". New methodologies are employed, such as using link shorteners to hide a reference to a disinformation website; and online abuse tactics, such as the off-platform coordination of attacks on minorities.



Threat actors use increasingly sophisticated tactics to avoid evolving detection mechanisms for online abuses. Image: Statista

Traditional methods aren’t enough

The basis of most harmful content detection methods is artificial intelligence (AI). This powerful technology relies on massive training sets to quickly identify violative behaviours at scale. Built on data sets of known online abuses in familiar languages means AI can detect known abuses in familiar languages, but it is less effective at detecting nuanced violations in languages it wasn't trained on – a gaping hole of which threat actors can take advantage.

While providing speed and scale, AI also lacks context: a critical component of trust and safety work. For example, robust AI models exist to detect nudity but few can

discern whether that nudity is part of a renaissance painting or a pornographic image. Similarly, most models can't decipher whether the knife featured in a video is being used to promote a butcher's equipment or a violent attack. This lack of context may lead to over-moderating, limiting free speech on online platforms; or under-moderating, which is a risk to user safety.

In contrast to AI, human moderators and subject-matter experts can detect nuanced online abuse and understand many languages and cultures. This precision, however, is limited by the analyst's specific area of expertise: a human moderator who is an expert in European white supremacy won't necessarily be able to recognize harmful content in India or misinformation narratives in Kenya. This limited focus means that for human moderators to be effective, they must be part of large, robust teams – a demanding effort for most technology companies.

The human element should also not be ignored. The thousands of moderators tasked with keeping abhorrent content offline must witness it themselves, placing them at high risk of mental illness and traumatic disorders. Beyond care for moderators, this situation may limit the operation's effectiveness, as high churn and staffing instabilities lead to low organizational stability and inevitable moderation mistakes.

The "Trust & Safety" intelligent solution to detect online abuse

While AI provides speed and scale and human moderators provide precision, their combined efforts are still not enough to proactively detect harm before it reaches platforms. To achieve proactivity, trust and safety teams must understand that abusive content doesn't start and stop on their platforms. Before reaching mainstream platforms, threat actors congregate in the darkest corners of the web to define new keywords, share URLs to resources and discuss new dissemination tactics at length. These secret places where terrorists, hate groups, child predators and disinformation agents freely communicate can provide a trove of information for teams seeking to keep their users safe.

The problem is that accessing this information is in no way scalable. Classic intelligence collection requires deep research, expertise, access and a fair amount of assimilation skills – human capacities that cannot be mimicked by a machine.

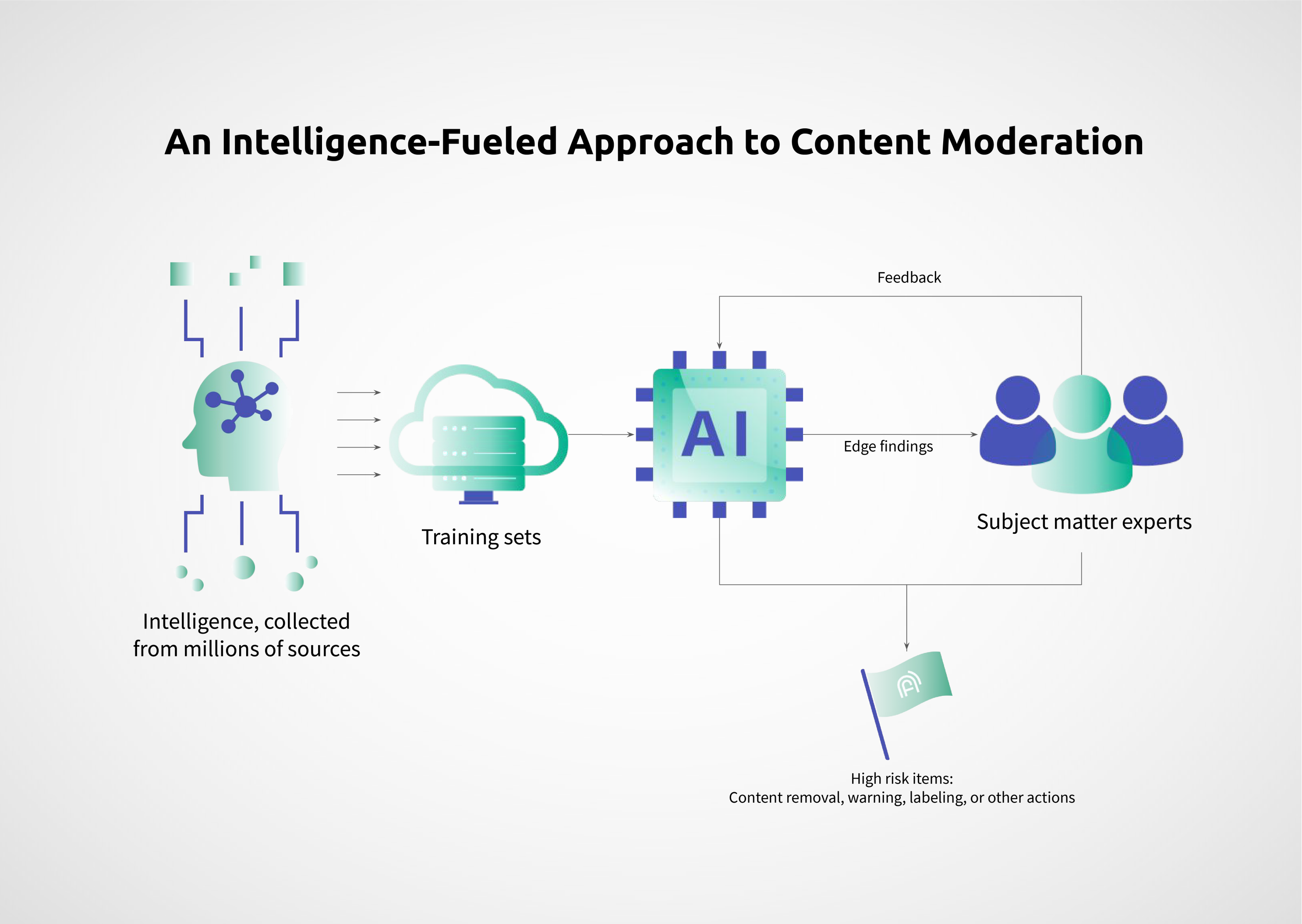
Baking in intelligence

We've established that the standard process of AI algorithms for scale and human moderators for precision doesn't adequately balance scale, novelty and nuance. We've also established that off-platform intelligence collecting can provide context and nuance, but not scale and speed.

To overcome the barriers of traditional detection methodologies, we propose a new framework: rather than relying on AI to detect at scale and humans to review edge cases, an intelligence-based approach is crucial.

By bringing human-curated, multi-language, off-platform intelligence into learning sets, AI will then be able to detect nuanced, novel online abuses at scale, before they reach mainstream platforms. Supplementing this smarter automated detection with

human expertise to review edge cases and identify false positives and negatives and then feeding those findings back into training sets will allow us to create AI with human intelligence baked in. This more intelligent AI gets more sophisticated with each moderation decision, eventually allowing near-perfect detection, at scale.


An Intelligence-Fueled Approach to Content Moderation
Intelligence, collected from millions of sources
Training sets
AI
Feedback
Edge findings
Subject matter experts
High risk items:
Content removal, warning, labeling, or other actions

By bringing human-curated, multi-language, off-platform intelligence into learning sets, AI will then be able to detect nuanced, novel online abuses at scale, before they reach mainstream platforms. Image: ActiveFence

The outcome

The lag between the advent of novel abuse tactics and when AI can detect them is what allows online abuse to proliferate. Incorporating intelligence into the content moderation process allows teams to significantly reduce the time between when new online abuse methods are introduced and when AI can detect them. In this way, trust and safety teams can stop threats rising online before they reach users.

License and Republishing

World Economic Forum articles may be republished in accordance with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Public License, and in accordance with our Terms of Use.

The views expressed in this article are those of the author alone and not the World Economic Forum.

Stay up to date:

Cybersecurity

Follow +

Related topics:

Cybersecurity Emerging Technologies Cybercrime

Share: