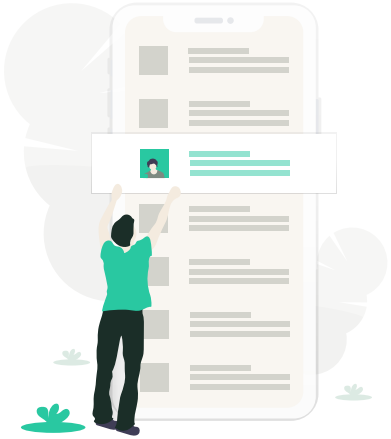
spectrum labs

How Social Media Platforms Can Build a Safe, Inclusive Environment

Assessing Trust & Safety

Summary

Trust & Safety professionals create a safe, trusted, welcoming environment on their social media platforms by managing user-generated content (UGC) - discouraging harmful content and encouraging helpful content.



80%

of all online content is user-generated

Whether you are just starting to introduce social features or UGC options to your community, launching a new social community, or looking to assess your current moderation efforts, this guide can help. It provides a simple, yet comprehensive framework for you to use in your efforts.

Every minute, users:

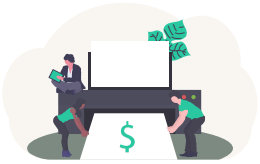
- Upload 300 hours of video to YouTube
- Post 317,000 status updates to Facebook
- Create 2.1M Snapchat photos
- Send 360,000 tweets

What is Online Toxicity?

The definition of what 'toxicity' means can vary – from one platform to the next, or for different audiences. Something completely appropriate and acceptable in one scenario can be exactly the opposite in another. Generally, however, these behaviors are considered toxic when they appear on social media:

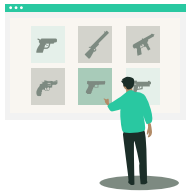
Child Sexual Abuse Material (CSAM)

Sex Solicitation, Misogyny



Scams, Fraud

Insults, Bullying, Harrassment, Hate Speech



Drugs, Weapons

Self-Harm

Graphic Violence

Terrorism, Radicalization

Effects of Online Toxicity

Toxic behaviors cause a variety of short-term and long-term challenges for individuals and platforms.

- Degradation of the intended user experience
- User and witness trauma
- Content moderator trauma
- Revenue loss from user attrition, ad revenue withdrawal, legal action

Degradation of the intended user experience

Product managers, engineers, and UI/UX designers spend hours designing products and features to enhance user experience and ultimately drive revenue. When people hijack products and features to harass or radicalize others, spew hate speech, or incite violence, the user experience that has been so carefully crafted is lost.

User and Witness Trauma

At the individual level, directly experiencing or witnessing toxic behaviors can cause stress, anxiety, and depression. Some people even suffer from PTSD after experiencing toxicity. Data labelers, moderators, and consumers alike suffer; and trust between users and platforms is degraded – damaging brand reputation, eroding user loyalty and causing churn, and ultimately affecting the bottom line.

Content Moderator & Data Labeler Trauma

As part of their job, content moderators and data labelers must review a variety of content: much of it illegal and violent, graphic and disturbing. This causes trauma in content moderators much like that experienced by healthcare and law enforcement professionals: leading to stress, anxiety, depression, and even PTSD.

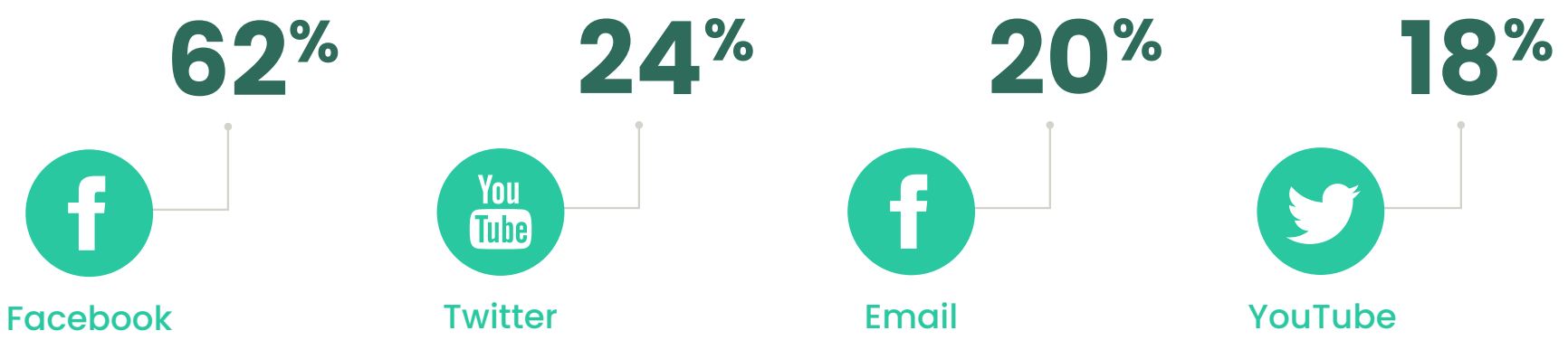
Revenue Loss

Trust & Safety managers are charged with the increasingly complicated mission: to ensure the intended user experience is upheld and that authentic interactions are supported; while community guidelines are also upheld and toxic behaviors are limited. Balancing these priorities can be difficult, but the end goal should be that the community interacts in the way it was intended. Toxic behaviors interrupt users, and negatively impact the user experience on your platform, which can ultimately lead to user churn and revenue loss.

The Effects of Toxic Behaviors

Toxic behaviors aren't just unpleasant in the moment, they can cause a variety of short-term and long-term challenges for individuals and platforms. At the individual level, toxic behaviors can cause stress, anxiety, and depression. Some people even suffer from PTSD afterward. Data labelers, moderators, and consumers alike are suffering; and trust between users and platforms is degraded – damaging brand reputation, eroding user loyalty and causing churn, and ultimately affecting the bottom line.

People Reporting Harassment:



Trust & Safety Framework

At Spectrum Labs, we work with some of the most innovative technology companies across dating, gaming, social, e-learning, and marketplace verticals. Our customers view content moderation as a vital activity key to creating great user experiences. Through our work with them, we've created a Trust & Safety Framework that helps everyone, from the most experienced Trust & Safety team to those just getting started, establish scope, and evaluate efforts.

The framework below demonstrates five distinct steps in developing and managing a trust & safety program, complete with questions that you should ask at each stage to measure and improve its effectiveness.

5 Stages of Trust & Safety

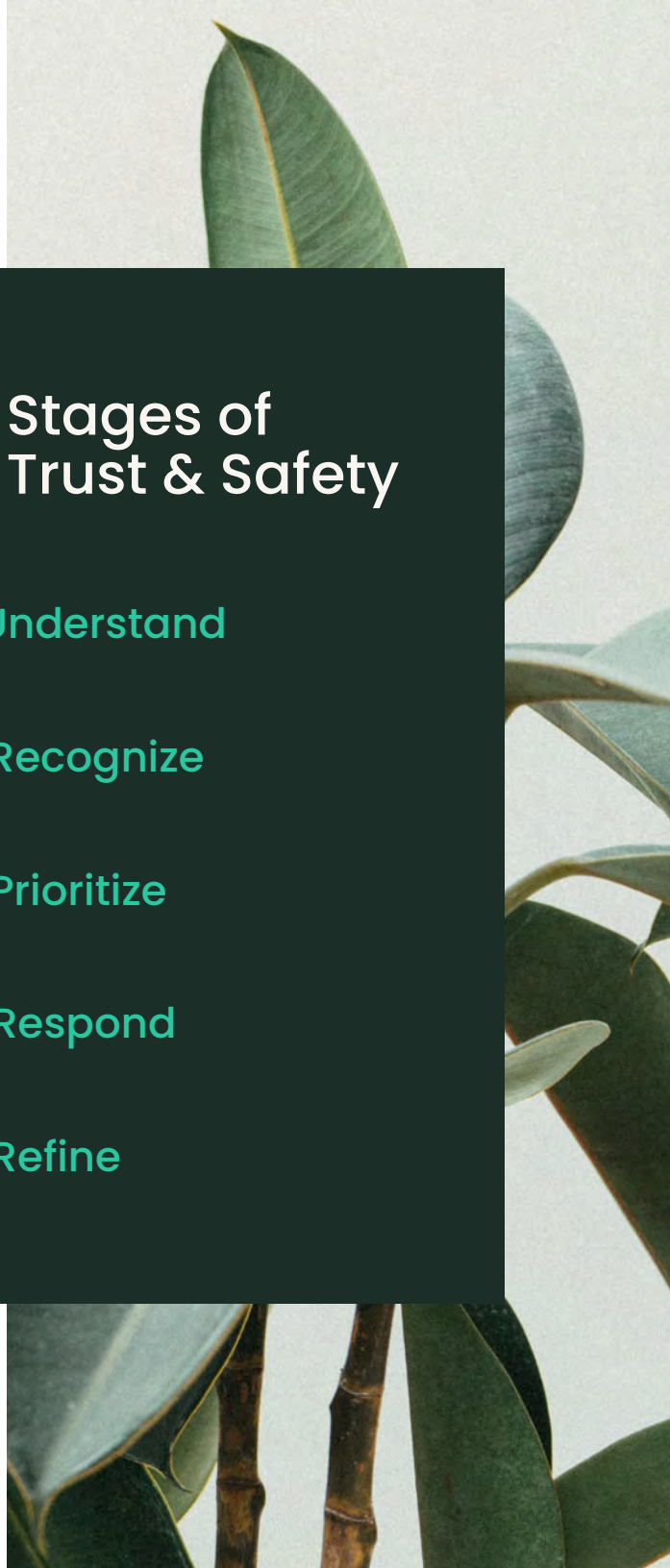
1 · Understand

2 · Recognize

3 · Prioritize

4 · Respond

5 · Refine



1

Stages of Trust & Safety

Understand

Do you know what is happening on your platform?

- Are you **staffed** as we described earlier?
- Do you have **tools that display behaviors** happening on your platform(s) in real-time and over time?
- Do you have prominently displayed **community guidelines** that clearly outline (and provide examples of) appropriate and inappropriate behaviors?
- Can they provide information gathered over time, for **tracking trends** and identifying changing tactics?

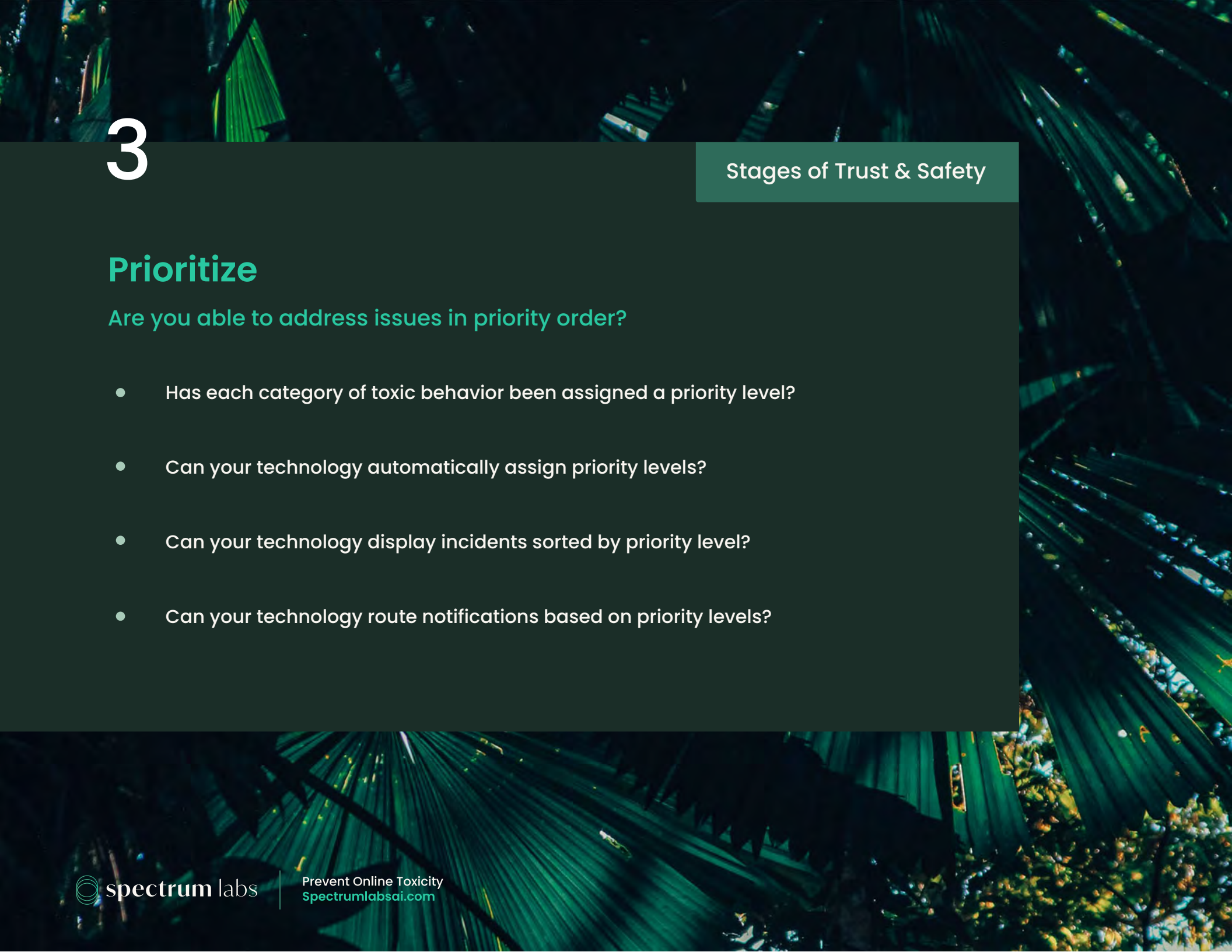
2

Stages of Trust & Safety

Recognize

Can your moderation tools identify toxicity accurately, in real time, and across languages?

- Do you have technology capable of detecting banned behaviors in real-time, across content types, communities and languages?
- Is that technology capable of applying different acceptance thresholds to different communities? (e.g. A community connecting teenage knitters should have different thresholds than one connecting sex toy enthusiasts.)
- Is that technology capable of self-learning to keep pace with your evolving community?
- Are instances of inappropriate behaviors added to an evaluation queue for review?



3

Prioritize

Are you able to address issues in priority order?

- Has each category of toxic behavior been assigned a priority level?
- Can your technology automatically assign priority levels?
- Can your technology display incidents sorted by priority level?
- Can your technology route notifications based on priority levels?

4

Stages of Trust & Safety

Respond

Are you able to apply nuanced responses?

- Do you have set consequences developed for each type of toxic behavior?
- Are multiple self-reporting features (block, flag, report, etc) available to users?
- Is your technology capable of executing consequences automatically in real-time?
- Is your technology capable of suggesting consequences, or displaying similar incidents and corresponding decisions for reference?
- Is your moderation team staffed to respond in a timely way to incidents?

5

Stages of Trust & Safety

Refine

Are you able to adjust quickly as new language and behavior norms arise?

- Does your team regularly update guidelines and deliver those updates to engineering and product?
- Is your detection technology self-learning?
- Do you have resources dedicated to regular detection updates to ensure you're catching the newest workarounds?

Content moderation is becoming more complicated. As the volume of user-generated content continues to grow exponentially and perpetrators evolve their approach to evade detection, many are struggling to find an effective Trust & Safety process and a supporting technological solution.

spectrum labs

Introducing Spectrum Labs

Content moderation is becoming more complicated. As the volume of user-generated content continues to grow exponentially and perpetrators evolve their approach to evade detection, many are struggling to find an effective Trust & Safety process and a supporting technological solution. Introducing Spectrum Labs

Spectrum Labs provides contextual AI, automation, and services to help consumer brands recognize and respond to toxic behavior. Our platform identifies 40+ behaviors across all languages enabling Trust & Safety teams to deal with harmful issues in real-time. Spectrum Labs' mission is to unite the power of data and community to rebuild trust in the Internet, making it a safer and more valuable place for all.

To see our solution for yourself, request a demo with one of our experts today.

sales@getspectrum.io
spectrumlabsai.com