How to Build a Trust and Safety Team In a Year: A Practical Guide From Lessons Learned (So Far) At Zoom

Karen Maxim, Josh Parecki, and Chanel Cornett

1 Introduction

"My boss asked me, 'do you have everything under control?' I was like 'yes' but I was thinking, what if I'm out sick? Basically, while I'm on vacation there's nobody there. It's just me; I'm it." -An early Zoom Trust and Safety professional, about their experiences in the spring of 2020

Trust and Safety (T&S) teams are most often born in a crisis. Based on our discussions with other companies, it seems rare that technology executives wake up one day and think, "Next quarter we should start a Trust and Safety Team." It's what you do when something bad has already happened. Maybe you notice a lot of cryptocurrency scammers contacting your users, or fake reviews directing users off the platform to hand over login credentials, or your app has become the gathering place du jour for a community of zoophiles. You grab whoever you can to address the problem immediately, and that's where T&S teams come from.

Zoom rapidly scaled and formalized its T&S team in the spring of 2020, at a time when we were growing at a blistering pace and gaining a larger, global user base at the same time. Suddenly, people were using Zoom in ways far beyond the business use it was designed for. Growing a T&S team almost from scratch, remotely, during a pandemic, has been one of the most intense professional experiences we've ever had. It has been exhilarating, exhausting, frustrating, fun, shocking and routine. And sometimes all of those in a single day.

We wrote this paper for those of you who may be joining or starting a T&S team at a tech company. If you're new to the world of T&S or building your first T&S team, this is for you. It is a practical guide, drawn from our early experiences and mistakes.

2 Hypergrowth and the Aftermath

Zoom goes from B2B to B2 everyone else. Before March 2020, Zoom had been designed for business customers. At the time, our abuse reports focused on a small set of violations of our Acceptable Use Policy (since replaced by our Community Standards), and the reports came from a handful of individuals in a handful of locations. Most of the reports involved what we call ICS meetings ("Illicit Content and Substances," or large sex and drug parties). Party organizers would repeatedly report each other in hopes of getting their competitors kicked off of Zoom. When Zoom took action against ICS hosts or meetings, the hosts and attendees would create new accounts and schedule new parties.

doi:10.54501/jots.v1i4.81

Then they would report each other all over again. It was repetitive, but also predictable and manageable.

As is the case at many small tech companies, Zoom's first T&S employees were in the customer support organization. Though there were a few additional employees in engineering and engineering support, T&S issues were primarily handled by a customer support specialist who addressed incoming abuse reports and a customer support leader who chipped in on more complex tasks. We were able to meet our users' needs.

The pandemic upended our processes. Zoom's growth exploded as people began to work, learn, socialize and do everything else from home. Many of our new users were individuals, broadening Zoom from largely business-to-business (B2B) into a business-to-consumer (B2C) company, too.

We began to get reports of an awful new phenomenon: meeting disruptions (a.k.a. meeting bombings). People started to disrupt meetings with conduct ranging from mildly annoying to heinous, hateful or illegal. In the early days, the disruptions were often students in remote classrooms — a digital version of pulling a fire alarm to get out of class. Over time, the disruptions got more organized, and the conduct became more serious and hateful. The trolls found each other and organized on various social media sites. They would exchange meeting credentials that had leaked or been posted on public forums. Then they used those credentials to disrupt the meetings (often in hordes) with malicious video, audio, annotation, profile pictures, screen names, backgrounds, chats, and using virtual cameras. The disruptors also targeted high-profile meetings for attention and the fun of causing pandemonium.

Suddenly we were inundated with reports of abuse from schools, churches, community groups and others who experienced these disruptions. It was difficult to watch the reports pour in faster than we could read them.

A few weeks after those early 2020 meeting disruptions, Zoom announced a freeze on new features for a 90-day all-hands-on-deck focus on privacy, safety and security.[1] Scaling our T&S function was one of the top priorities. Zoom needed more people, more resources and new tools.

Here is a partial list of the higher priority tasks for the T&S team as we rapidly scaled to meet an explosion of demand:

- Promote and explain existing safety tools
- Develop new safety tools to support all the new uses of Zoom
- Create backend infrastructure to respond to and act on reports of abuse
- Build data pipelines and a dashboard to enable transparency reporting
- Build a policy framework and internal processes to guide T&S actions
- Develop relationships with customers, civil society groups and governments to get feedback and input
- Publish a government requests guide
- Address a crush of internal and external questions, complaints, concerns, escalations, legal actions and investigations

In May 2020, Zoom made its T&S function a standalone team and hired an experienced lawyer to lead and build it. There was barely time to run a hiring process in those early days, so we recruited smart people from around the company to build our ranks. We

[1] https://blog.zoom.us/a-message-to-our-users/

pulled in a program manager from the Marketing team, a generalist attorney who was serving a few different teams, analysts from Zoom's Support function, engineers, and others. Most of these people did not have previous T&S experience. As soon as we began hiring externally, we brought in team members with T&S expertise. Two years later, we have a core T&S team of 27 and a cross-functional team of about 35 additional subject matter experts (e.g., T&S Engineering).

Develop new safety tools to support all the new uses of Zoom. As the hypergrowth started, T&S was accepting and working on reports from customers via a flat email-type system. Users would send an email to a T&S alias account. T&S would then work through the cases as they came in, searching for the most urgent ones. The team also responded to ad hoc escalations from across the company. Email reports rarely included sufficient information for T&S analysts to investigate, so the analysts would have to correspond with individual customers, who were often upset. Those interactions frustrated our users, protracted each investigation and drove caseloads up further. Our small team of analysts was working around the clock to try and catch up, but the system was not designed to scale to our new millions of users. Our first-come-first-serve email queue favored the oldest cases, the most insistent voices and those who escalated through other Zoom employees, rather than the most serious cases. Our queue was growing faster than our analysts could clear it, they couldn't prioritize all the urgent requests, and they were exhausted and dejected.

There were three critical areas we needed to build up quickly:

Table 1: Overview of the three critical areas of Trust & Safety at Zoom

Frontend	We use "frontend" to refer to the part of the product the user interacts with. The frontend tools we created allow the host to: • Control who enters a meeting • Determine who uses video, audio, chat, and other modes of communication in a meeting • Remove individual unwanted or disruptive participants within a meeting • "Suspend Participant Activities" or "pause" a meeting to remove and report an offending party and prevent further disruption • Report an abusive user from within the meeting • Report abuse outside of a meeting • Receive automated notifications of specific adverse actions along with a link to an appeal form • Appeal adverse actions taken against their accounts • Receive notification from our "At Risk Meeting Notifier" that looks for public posts with Zoom meeting links that may be at risk for disruption. We send a warning and suggested steps for securing the meeting

Middle	“The middle” is the part of the product where data is collected, protected, processed and transferred. We: • **For reports about abuse in meetings:** Automatically pull and preserve all relevant metadata, capture it in defined tables and then pass it to our backend enforcement dashboards • **For reports of abuse outside of meetings:** Do the same as above. Non-meeting abuse like account takeovers, fraud and spam involve different metadata. That information needs to be linked to the reports and pushed to backend dashboards so our analysts can investigate and act efficiently • **For appeals:** Securely link the appeal to the underlying action • **For notifications:** Assign codes to each Term and Standard. We use those codes to automatically send particularized notification to users based on violation type and enforcement action • **For prioritization:** Developed automated prioritization based on abuse report type, in order to handle the most serious matters first • **For transparency reports:** Map our abuse reporting interface and our law enforcement requests interface to collect, de-identify, categorize and organize information for our two transparency reports • **For learning:** Generate and collect meaningful data across T&S infrastructure for us to learn from and refine our processes
Backend	“The backend” is where humans or machines implement T&S policies. We have developed: • An analyst dashboard that prioritizes abuse reports by severity, and displays reports and relevant metadata in one place. The dashboard offers analysts a menu of options to quickly and efficiently take appropriate action on each abuse report and appeal • API integration with the National Center for Missing and Exploited Children to report child sexual abuse material and with the FBI’s National Threat Operations Center for reports involving threats to life and physical harm • Account theft protection tools that identify Zoom users whose login credentials may have been stolen or compromised in a data breach elsewhere on the internet. Our tools notify them and prompt them to reset their password • Heuristic models and machine learning to analyze data associated with abuse reports for more proactive detection and prevention of abuse

Starting at the frontend, the first task was to build additional safety features into the product to arm Zoom users with tools to better protect themselves and their meetings. These include features such as “Suspend Participant Activities,” the “At Risk Meeting Notifier,” the ability to limit attendees to specific regions/countries, and better abuse reporting forms, both from within the meeting and outside of it.

Next, we needed to improve the ways that users reported abuse, both from within a meeting and after it ended. In the case of in-meeting reporting, it has to be intuitive enough for a panicked user to quickly find in the stress of a disruption, but not so easy that people can create a lot of false reports or otherwise abuse the feature itself. Users can now submit in-meeting reports via the “Participant” or “Meeting Information” icons, which are prominent and easy to find. We also created a Trust Form for users to submit reports after meetings or to report abuse outside of meetings.[2] In collaboration with our Communications team, we make it a priority to educate our users on how to use these tools via blogs and support articles.[3]

In the middle, we had to begin collecting the right information, and enough of it, to securely generate cases for analysts to evaluate without back-and-forth communications. Our tools needed to categorize the reports, order them by urgency so we could prioritize the most serious ones (e.g., those relating to child sexual abuse material), and generate data for transparency reports. Zoom now maintains two transparency reports: one for T&S enforcement actions and one for law enforcement requests.[4]

On the backend, we built an access-controlled, secure database and dashboard for analysts to review reports, take action and automatically report cases involving child sexual abuse material or imminent threats of violence or self-harm to the right authorities. We nicknamed this dashboard “One Page, One Click” or OPOC. OPOC has evolved a lot as Zoom’s features have changed and the name is increasingly inapt. But we continue to strive for simplicity and streamlined workflows. Our analysts should have all the information they need to investigate a report on one page, and be able to investigate and enforce on the same page, with the fewest number of actions.

Rebuilding the frontend, middle and backend tools from a T&S perspective was a sea change for our team. As the new tools came online, analysts moved from the edge of burnout to becoming energized and “crushing queues” that once seemed overwhelming. That freed up analyst time to more actively contribute to product and engineering designs, including new functionality such as automatic notifications, appeals, proactive detection and prevention of abusive conduct.

3 Nuts and Bolts

3.1 Get Involved Early

To be effective, T&S has to be involved early and often throughout the complete product ecosystem. Building a safe, privacy-preserving product ecosystem requires T&S to touch the frontend, middle and backend of the product and platform.

Here’s an example of why T&S should be involved early. Imagine that your Product team announces that starting next week, customers will be able to leave reviews of the widgets you sell on your website. As soon as the review function launches, disgruntled customers (and potential trolls?) begin leaving nasty, profanity- and hate-laden reviews on your

[2] https://zoom.us/trust-form

[3] https://blog.zoom.us/safer-internet-day-2022/; https://support.zoom.us/hc/en-us/articles/360042791091

[4] https://explore.zoom.us/en/trust/community-standards-enforcement/; https://explore.zoom.us/docs/en-us/trust/transparency.html

beautifully-designed website. There's no way to confirm whether a reviewer actually bought a widget. There's also no way to remove an offensive review or block people who try to leave dozens of such reviews. Nobody can make principled, consistent decisions about what reviews to take down, because nobody wrote any rules. In our experience and from talking with others in the field, as soon as you launch your brand new product or service into the world, someone, somewhere will start trying to abuse it for fun or profit.

If your hypothetical Product team had involved T&S as soon as it began designing the review feature, T&S would have been able to spot most of the avenues for abuse: It's T&S's job to predict and prevent abuse before it happens. T&S can help develop ways to get the benefits of posted reviews while minimizing the risk of bad behavior. When T&S is not included at the outset, everyone responsible for the new feature has to scramble when it launches and the abuse begins. It's enormously stressful and a bad look for all involved.

It's also T&S's responsibility to help company decision-makers understand the cost of new features. When the Product team comes to you with, say, plans to develop a comments feed in a new feature, you will need to explain that if the comments are to be free of bots trying to get you to "mint a Roostr" in the chicken metaverse, you will need enough lead time to build the anti-abuse tooling, write the rules of the road, and hire analysts to review reports.[5]

T&S in general, and content moderation in particular, is expensive. That cost needs to be accounted for in any plan to expose the company to user generated content (UGC).

3.2 Write the Rules

When we arrived in spring 2020, Zoom's Terms of Service (TOS) and Acceptable Use Policy (AUP) governed the entire platform. Those documents included basic rules like "Don't do illegal stuff," "No spam or malware," "Don't use this product to operate nuclear facilities" (full disclosure, that last one is from Keybase's AUP; Zoom acquired Keybase in May 2020). But you may need some extra guidelines to govern user interactions and content if you have things like:

- Publicly accessible UGC such as reviews, posts or profiles
- Places where your users interact with each other

Most often, these guidelines live in an AUP, Community Guidelines or Community Standards (CS). A CS or AUP are more definitive than a TOS and are designed to give a customer a more transparent understanding of how a platform views content and the enforcement processes.

Imagine, for example, that an adult video news outlet planned to host its annual awards ceremony on Zoom Events. Based on the past awards ceremonies, we can be certain that this event will feature lots of sexually explicit clips. What should Zoom cite when it disallows the event?

[5] https://chikn.farm/

Table 2: Relevant extracts from Zoom's Terms of Service and Community Standards

Zoom Terms of Service	Zoom Community Standards
You may not...use the Services to communicate any message or material that is harassing, libelous, threatening, obscene, indecent, would violate the intellectual property rights of any party or is otherwise unlawful, that would give rise to civil liability, or that constitutes or encourages conduct that could constitute a criminal offense, under any applicable law or regulation *(emphasis added)*	[You may not use Zoom for] adult content [, which] is any media that is pornographic or intended to be sexually gratifying, whether photo or video, cartoon or animated.

Zoom's TOS, like most online terms, include broad prohibitions; one could use the bolded clause alone to act against virtually any kind of abuse on Zoom. Both sets of rules allow Zoom to act, but we find decisions to be clearer and easier to explain when we rely on the more, uh, explicit, ones.

We wrote our first set of Community Standards because we were about to release a new product called OnZoom. OnZoom is a place to host events for the public, like classes, workshops or fundraisers. It has a public-facing directory of events that anyone can browse. It was Zoom's first foray into public-facing UGC, and it meant that we needed a set of rules that would clearly explain what was and was not acceptable for those events.

We started with Zoom's values and core principles — these are our North Star. Over a series of meetings, we wrote down a set of sentences that tried to capture those values in the context of a set of rules for content at Zoom. Those principles are the basis of the preamble and "Zoom Content Moderations Principles" at the top of our CS.

Then we benchmarked extensively. We read the rules of companies with products similar to ours and other well-established companies. The more mature a company is, the more likely it is that its rules are well-vetted. Another source of language for platform rules is civil society organizations. Tech Against Terrorism, for instance, has some excellent model policies. A practical tip: use spreadsheets to collect and compare different clauses.

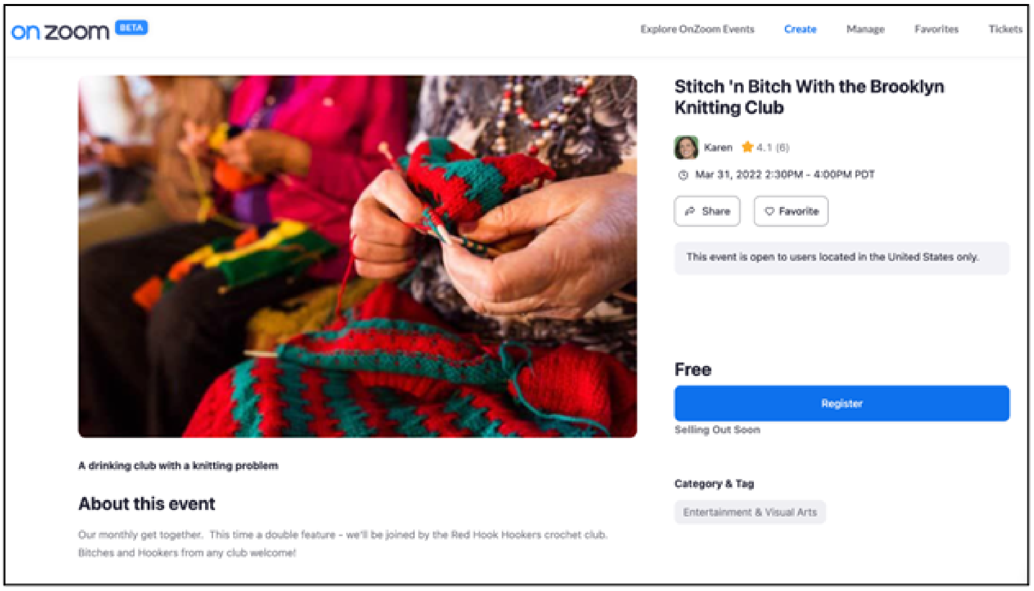
We once got some great advice about writing our first set of rules from a professor in the field. Moderation, she said, is a one-way ratchet, so start with lighter rules and add more as needed. Once a rule is in place, it's hard to walk it back. You can see this in the evolution of rulesets at just about any platform. The rules almost always get longer and more specific over time.

Another reason to start with lighter rules — a principles-based approach over a rules-based approach — is that you can't anticipate all the novel issues you'll need to address over time. You will need to leave some flexibility in your rules so that you can address new issues without having to add to or change the rules every time. You can spot flexibility by looking for phrases like "may", "such as", "including", or "as appropriate". But if you have too much of that soft language, you will feel arbitrary when you make decisions and look that way to others.

3.3 Get Input and Approval

The more people who read the draft rules the better. You'll want buy-in from key business leaders in addition to Legal, Government Relations, Communications, Privacy and Compliance, for example. These other teams can spot the ways that your draft rules will interact with existing laws and regulations and ensure that the rules are consistent with your company's broader priorities and goals.

During the course of creating the CS, we workshopped with the executive team to align on our core principles for our first set of rules. This is important: Your core principles should be endorsed at the top. We prepared slides advertising fictional events that played in the gray areas of our draft rules, using the actual OnZoom user interface (UI).



Stitch 'n Bitch With the Brooklyn Knitting Club

Karen ★ 4.1 (6)

Mar 31, 2022 2:30PM - 4:00PM PDT

This event is open to users located in the United States only.

Free

Selling Out Soon

A drinking club with a knitting problem

About this event

Our monthly get together. This time a double feature - we'll be joined by the Red Hook Hookers crochet club. Bitches and Hookers from any club welcome!

Category & Tag

Entertainment & Visual Arts

Figure 1: The kind of fake event we presented for the executive team to align on standards around profanity

We showed each slide and then — without discussion, so that people wouldn't influence each other — we polled everyone about what they would do about each meeting. (See for example Figure 1.) After all the polling, we shared the results and only *then* started the discussion. Later, we made a second set of slides with new scenarios around the most disputed (or "edge") cases. We met again with the executive team and ran the same exercise with the new slides. After that second meeting, we finalized and published the CS.

We felt it was important to present the fake events in our new UI, to allow the Executive team to experience their own gut reactions. Product and Executive teams are used to seeing marketing materials with glossy stock photos of beautiful people in yoga or cooking classes. It's hard to explain the potential abusive or ugly ways a product might be used, and all the edge cases. But it's easy to *show*. We wanted our Executive team to experience the full range of content that we could expect to see when OnZoom launched.

The other thing we hoped to achieve in the meetings was an express endorsement of our proposed approach to T&S. We sought for the T&S team to be empowered to create and follow processes to decide hard cases, and for Zoom's leadership to back us up both internally and externally when needed. In the beginning of a company's T&S journey, controversial T&S decisions are usually made by one or more executives. It is a *big deal* to entrust those decisions to anyone else. It can take time and a record of good judgment to earn the trust and autonomy you need to run an effective T&S program.

3.4 Process, Process, Process

Good process goes hand in hand with good rules, and is just as important. No matter how good your T&S rules are, there will be difficult cases that fall into gray areas. Without solid processes for making decisions, each hard call can become a chaotic fire drill. Anyone with an opinion or a stake in the outcome may try to get involved, and that will be overwhelming. Or, equally bad, *nobody* wants to make the decision, so it devolves to a single lonely person (and potential future scapegoat).

There are several other reasons to strive for great T&S processes. First, you can lean on them when it's not intuitive what to do, which will be often. It's comforting to know, "This is a tricky issue, so I'm going to follow the flow chart we made exactly for these kinds of cases." It helps the analysts have confidence in their decisions.

Second, good process protects your decisions. You never want to find yourself defending a tough call by trying to remember why you made the decision or explaining your intuitions. It is much better to say, "I followed the process in our playbook."

Good process helps you build trust with your leadership team. The more you use your processes and rely on them successfully, the more your leadership will come to trust them and you, too.

Keep in mind the old saying: are your processes *repeatable*, *scaleable* and *auditable*?

- *Repeatable:* so your team can carry out the same process the same way in different instances. Once you mature, consider creating a Quality Assurance program to spot check your decisions for consistency
- *Scalable:* as your company grows, your processes will need to cover more and more cases, so they should be low-touch and easy to apply
- *Auditable:* keep records of all your decisions in an easy to find, searchable format. This helps you maintain the truth and fairness of your decisions and allows you to set or revise precedents accurately

All of our process documents live in our internal T&S playbook and we review them on a schedule (following a process, of course!). We use branded headers and Zoom's official logo for each policy, so they appear as official to others as they are to us. That may sound trivial, but if a process *looks* official, it's more likely to be taken seriously. Appearance matters.

A word about version control: We date each process at the top, and once the date is on it, we aim not to touch the document again. When a process needs an update, we create a brand new document and move the old one to an archive. Also, we strive to write processes so that anyone with the appropriate level of access can grab the process and execute it, even if they've never used the relevant system or encountered the type of issue before. Each process must be easy to understand and fast to execute.

Your processes will vary widely based on your revenue model, the expectations of your users and where in "the stack" you are. Your level in the stack depends on whether you're an internet service provider, cloud service, app, platform, content delivery network, ethernet cable, etc.[6]

Zoom's products were designed to function as a private conference room or living room and our internal processes reflect that. The vast majority of our meetings are small and intimate; we do not monitor live meeting content, and our users expect that we will not. Our processes reflect that Zoom does not have access to meeting content unless a user chooses to share it with us as a screenshot, description or recording. For example, in most of the reports we receive there is no documentary evidence. That's because most of our users' content is as ephemeral as a face-to-face conversation. We rely on metadata and use a U.S.-legal-system "preponderance of the evidence" standard to make decisions, which means that in many cases, the report is dismissed for lack of enough evidence to take action.[7] By contrast, a social media platform or a retailer may have access to the content in question, which persists in feeds, messages between users, or in posted reviews.

As of this writing, Zoom T&S has many different policies, processes, memos, flowcharts and white papers that govern our work (see the Appendix for some examples); in fact there might be more of these documents than there are T&S people at Zoom! Perhaps that ratio will flip someday, but for now it simply reflects that we're trying, very carefully, to build for the future.

3.5 Who Decides?

There are a lot of options for who decides T&S matters. In general, we have found it best to spread the decision-making among multiple people. This issue came up in the Arbiters of Truth podcast episode "Content Moderation's Original 'Decider.'"[8] If one person is the main decider of the hardest T&S questions, the company is vulnerable to accusations of arbitrariness. And if only one person makes the hard calls, that will be who's summoned by the government to explain controversial decisions. It's much better if the T&S lead can describe a straightforward process and point to a body of decision-makers, instead of having to describe how they alone made the decision. Finally, being "The Decider" gets to be too much work for one person, particularly if your company operates in multiple time zones and jurisdictions.

Zoom T&S decisions are made in a tiered review system consisting of four levels of review.

Tier I: Our Tier I reviewers are trained on Zoom's platform rules, have passed a resiliency screening, and have mental health resources available to them. They are a combination of Zoom analysts and employees of a contractor that specializes in T&S. Tier I reviews reports flagged or submitted by people or automated tools for alleged violations of the CS. If a Tier I reviewer cannot make a decision quickly, either alone or in consultation with a peer or supervisor, then they escalate the ticket to Tier II.

Tier II: Tier II analysts are employed by Zoom and are members of the U.S.-based T&S team. They review Tier I escalations and also do the initial review for some categories of alleged violations. For example, Tier II does the first review of all copyright-related tickets, because DMCA is a U.S. law that requires specialized training. Decisions that Tier II reviewers cannot make quickly are escalated to Tier III.

[6] For an explanation of content moderation in the tech stack, see Joan Donovan's "Navigating the Tech Stack: When, Where and How Should We Moderate Content?" here: https://www.cigionline.org/articles/navigating-tech-stack-when-where-and-how-should-we-moderate-content/

[7] For stats on enforcement actions, see our Community Standards Enforcement Report here: https://explore.zoom.us/en/trust/community-standards-enforcement/

[8] https://www.lawfareblog.com/lawfare-podcast-content-moderations-original-decider

Tier III: This is the T&S leadership team. In addition to reviewing reports escalated from Tier II, Tier III reviews controversial, hard-to-classify reports. They also review cases that come from elsewhere in the organization, such as the Communications or Executive teams. Tier III makes decisions on a consensus basis and memorializes all decisions in a decision bank. If Tier III cannot reach a consensus, they escalate to Tier IV. Tier III always errs on the side of escalating, especially when the case involves a new issue.

Tier IV: Tier IV is the Appeals Panel. It is a group of eight Zoom employees nominated by the Zoom senior leadership team. Tier IV decisions have a unique gravity to them because of the way Tier IV is organized and selected. Here’s how we describe the panel when we solicit nominees:

What makes for a good set of nominees? As a group, they should be:

- *Open-minded and willing to hear a broad set of perspectives in problem-solving*
- *From a diversity of teams across Zoom*
- *From a diversity of backgrounds, experience levels and tenures at Zoom*
- *Effective listeners*
- *Demonstrated track record of making decisions with good judgment*
- *Willing to look at and talk about things that make some people uncomfortable, like hate speech and nudity.*

Once the nominees are selected and they confirm they’re willing to serve, we hold a ceremony to administer the oath and swear them in. The Panel meets monthly, whether or not there’s a case to be heard, so T&S leadership can update them about trends in the lower tiers. Panelists serve for one-year terms, but can serve for up to two years. At the end of a term, we ask for four volunteers to leave the panel, so that each panel has a mix of new and experienced members.

3.6 Hire (and Borrow) a Good Mix of People

If you are creating a T&S team in a hurry, some of your first team members may come from elsewhere within your organization while you work to scale up. Pre-pandemic, our main T&S person had extensive T&S experience at another tech company, but others did not. In June 2020, we hired two analysts from outside of Zoom who both had prior T&S experience. The remainder of the early crew were existing Zoom employees. While not from a T&S background, our internal hires brought experience with Zoom’s existing internal tools plus essential contacts for us elsewhere at Zoom. Bottom line, over half of the first T&S members did not have prior T&S experience.

Since those early days, most of our hires have had prior T&S experience. Hiring people who’ve worked in the field at other companies, and at a variety of companies, provides significant value; these people have learned important lessons and gained perspective from their prior T&S work. To have a mix of people with prior T&S experience and people who are experts in your product and tooling is a powerful combination.

When we hire, the qualities most important to us as a T&S team are comfort with discomfort, the ability to work with “gray area” issues, and a general curiosity about the world.

Most of the people in T&S will at some point have occasion to see hate imagery or nudity in their work. We have measures in place to soften the impact of those encounters (such as image blurring, rotating personnel, and more) and prioritize our team’s mental health needs, but we also need to hire people who won't run away when confronted with those things. In interviews, it's important to speak candidly about the kinds of unpleasant things your team sees so that your new hire isn't surprised later.

We also have interview questions to test applicants' comfort working in gray areas. On a number of subjects, Zoom's rules for the platform are deliberately flexible so that we can take context into account (nudity and profanity are two such topics). We want to make sure that when a T&S person reviews an image that might violate our rules, the fact of the image is the start of the conversation, not the end of it. Specifically, what is the context? Is it nudity in a medical scenario? A breastfeeding group? Is violent or extremist content in the context of journalism?

3.7 Organize the Team

There is plenty of guidance on how to organize teams generally, and we won't try to duplicate it. So here's a very brief overview of how our team grew and organized over time.

In the beginning, everyone did everything. We had a flat hierarchy, with the Head of T&S at the top and everyone else underneath.

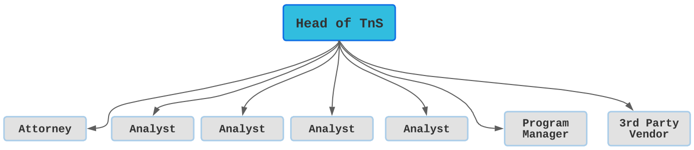


Figure 2: Zoom Trust & Safety's initial organizational structure

Over the past two years, it made sense for us to begin specializing. The way we've segmented our work reflects the changes in Zoom's needs over time. Our largest team is the core team that responds to user reports (also known as our "Protection and Recovery Team"). We added teams for in-depth investigations, specific types of fraud, and policy. We have several dedicated program managers who tackle long-term projects, plus a lead for quality assurance and training, who helps maintain consistency throughout our work. Around two years in, T&S looks like this:

Not everyone who works on T&S needs to be in the T&S reporting chain. The T&S engineers — most of whom work exclusively on T&S matters — report into the Engineering organization.

Zoom T&S is located within the Compliance function, which reports to the Chief Compliance, Ethics & Privacy Officer and, in turn, to the Chief Operating Officer. But T&S can be located in lots of different places within a company. For example, if most of the abuse you see is technical or financial (e.g., credential stuffing attacks or credit card fraud), the best place for you might be within an Engineering or Finance team, with strong support from Legal. Other T&S teams start out in the support organization. In practice, we suspect that T&S teams tend to be located — at least at first — wherever the first people working on T&S matters happen to have been.

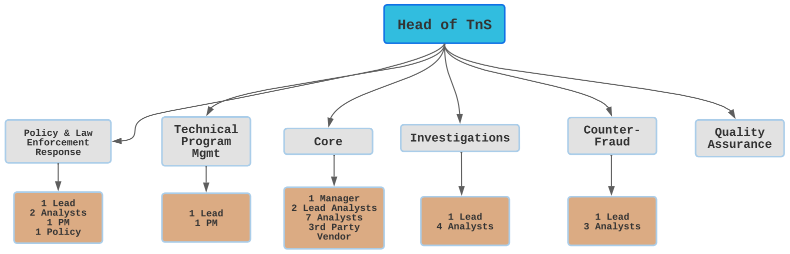


Figure 3: Zoom Trust & Safety's current organizational structure

Another consideration for where a T&S team lives is whether it has responsibility for law enforcement response. Law enforcement response can require specialized legal expertise and could counsel for your T&S team living under Compliance, Legal, or C&E.

To the extent that you are able to influence your organizational structure, you may simply want to aim for an effective leader, or one who deeply understands the work you do, or who will give you the most autonomy to work effectively, or the one closest to the company's important decision-making. Do what makes the most sense for your work and the organization's needs.

4 Make Friends

4.1 Inside the organization

We've found it critically important to develop working relationships with a variety of other teams within Zoom. We work closely with the Government Relations team, since both law enforcement response issues and decisions around abuse can have big implications for our operations in other jurisdictions. We also work closely with Litigation, Communications, Privacy, Compliance and Ethics, Product and Engineering.

We maintain regular meetings with contacts in those teams and include them on any matters we come across where there's even a remote chance that it involves their remit. In a couple of instances, we have written procedures for how our teams will work together, most notably, between T&S and Communications.

We have a guest speaker program in our T&S all-hands meetings. Anyone in T&S can invite anyone else from Zoom and beyond, and we interview the guest, podcast-style, about their work and how we can best collaborate. These interviews help us build relationships with groups we don't regularly work with and help us to put faces to names. That's especially important in Zoom's fully-remote environment.

We are also building a voluntary rotation program. Our hope is that T&S team members will gain new experience both within the team – by trying out work in our different specialties – and elsewhere in Zoom. Our vision is to continue building relationships within Zoom's other teams, and to keep our own people engaged and learning new things.

Finally, do your homework. You need to understand your products and systems at a deep level—as technical as you can possibly get. You need to understand how the engineers and product people work and how their teams are structured. The more you understand and speak their language, the quicker you can get things done.

4.2 With T&S teams elsewhere

Reach out as often as you can to any organization with products or problems similar to yours. The people who work at civil society organizations know everyone in the business and are great connectors, as are vendors that serve T&S teams and T&S professional organizations (see Section 4.3). If you can't find someone to introduce you, you can always try cold emails through LinkedIn or the good old guess-someone's-corporate-email approach. We've found that almost everyone has been willing to talk to us and we're grateful for their advice. When you're just starting to build, it's comforting to learn that you've managed to independently devise an approach that's industry standard (or at least not totally wacky). Your peers will give you good ideas, help you know whether you're on the right track, connect you with others and introduce you to the right organizations to join.

4.3 Professional organizations, civil society groups and other coalitions

Depending on your budget and size, you will want to join some of the groups and organizations that support T&S work:

Professional organizations: The main ones at the moment are the Trust and Safety Professionals Association (TSPA) and the Digital Trust and Safety Partnership (DTSP). TSPA is an excellent resource for individual professionals working in T&S; it's a safe space for us to support each other and discuss common issues in our work. DTSP focuses more on industry-wide policy issues, though both organizations work on policy.

Civil society groups: These are non-governmental organizations that work on specific issues, such as fighting terrorist content or child sexual abuse material online. They are a great resource for T&S teams. They can help you refine your rulesets or create good processes. They often have libraries of content like model policies and guidance on transparency reporting. Some of them keep track of new technologies that can assist your work. And they offer another space for you to connect with peers at other T&S teams and think through tough issues. One of the first organizations we joined is the Global Network Initiative (GNI). GNI provides a critical setting for companies such as Zoom to share developments in their T&S journey, particularly through the lens of freedom of expression and privacy. GNI participants from civil society, academia, and the private sector provide invaluable feedback on how member companies can improve their approach to T&S and prepare for high-profile events where their products could be used in harmful ways.

Coalitions and other groups: These are usually based in Washington, DC and work on issue- or sector-specific policy. There are a number that are relevant to T&S, such as Reform Government Surveillance, I2 Coalition and the Internet Society.

All of these organizations are funded by their members, and they will ask you for financial support, either as a membership fee or as a voluntary contribution. The fees are usually on a sliding scale based on the size of your program or your company's market cap. They may also be negotiable.

5 Get Educated

Because Zoom operates globally, it is critical that some of the people on our team know what's going on in geopolitics, and that we get our information from high-quality sources that publish reporting, not just commentary. The most difficult content-related cases we see usually involve sensitive geopolitical or cultural matters, and we need high-quality information to make high-quality decisions. Collectively, we like *The Economist*, *The Wall Street Journal*, *The New York Times*, *The Washington Post*, *The Information* and *Foreign Policy*. A couple of other teams within Zoom send regular summaries of news issues that could affect the company. Every so often, we need an in-depth briefing about a particular matter; two different teams assist us with those.

We also recently started a "reader" program. One of our attorneys compiles a biweekly newsletter for the team with current events updates that could impact our work. A typical newsletter from early 2022 had alerts about high-profile meeting disruptions, the 2022 winter Olympics, Russia/Ukraine developments, and a list of talks, conferences, and articles about T&S work. We recently began publishing our newsletters for all Zoom employees, in case anyone else is interested in the state of the field.

We keep a wiki for the team of interesting articles, books and podcasts that relate to T&S. Below are a few of the many excellent sources of information and news about T&S, digital speech, and tech policy issues—these sources are good for people who are brand new to T&S work.

Some Writing We Found Helpful

- Your Speech, Their Rules: Meet the People Who Guard the Internet, Alex Feerst
- The New Governors, Kate Klonick
- The TSPA Resource Library
- The Twenty-Six Words That Created the Internet, Jeff Koseff
- Three Eras of Digital Governance, Jonathan Zittrain

Podcasts

- Arbiters of Truth
- Pivot
- The Sunday Show (from Tech Policy Press)

Blogs/Newsletters/Sites

- Platformer
- The Lawfare Blog
- Everything In Moderation

Academia

A few of us have become major enthusiasts about the academic work being done on T&S matters. One great thing about working in T&S is that these brilliant thinkers will usually answer our emails. We need them for ideas and benchmarking; they need us for information about what's happening on the ground. When we started out, we hoped that these learned scholars would have the solutions to our hardest T&S problems. To our dismay, they usually do not. But we learn a lot from their work, and maybe someday they'll tell us the right answer. Also, there are a number of centers within universities that focus on internet governance issues, and we recommend following them for updates and events:

- The Knight First Amendment Center
- The Berkman Klein Center
- The Center for Internet and Society
- The Stanford Internet Observatory

6 Conclusion

It must be a universal experience for a new T&S person to feel unqualified on a daily basis. T&S as a field is young enough that many of its professionals have not had (much) prior experience in it. It's possible you don't either, like half of our original team. Soon you will face your first perplexing cases. Does a Civil War reenactment violate your weapons policy? How do your rules apply to a pole dancing fitness class? Is that a nipple? You may wonder why you're qualified to make these kinds of decisions. Hopefully you do feel that way, in fact, because it means you have the humility necessary to do good work in this field.

Those of us who focus on legal and policy matters feel especially humbled. The work we are privileged to do is meaningful and has real impact on the lives of our users. All of us who work in T&S owe it to our users to stay informed, learn from each other, acknowledge mistakes, scrap policies that don't work, and to keep innovating as the needs of our users change.

Luckily, the people who work in T&S are thoughtful, innovative and intensely pragmatic. They've created transparency reports and guides to transparency reporting. They developed PhotoDNA and ways to detect abusive behavior using only metadata. They have created tools and processes to decide speech matters at scale and offer appeals. They've invented warnings, labels, strikes, quarantines, geofencing, demonetization and other ways to bring proportionality to enforcement that once only had two tools: leave it up, or take it down.

The world of online speech is in flux. The most difficult speech quandaries no longer unfold in shopping malls, town squares or in newspapers. They happen on the proliferation of private online platforms, and the decisions are made within the corporations that maintain them. Governments around the world are advancing ways to shift the hard decision-making (or at least oversight of it) to their own policymakers, but so far their efforts are mostly in progress, just beginning to be tested, or otherwise unsettled. Even the trusty old First Amendment – whose precedents for content moderation have been settled for a while now – may be on the verge of a judicial transformation.[9] You can take comfort in the fact that nobody knows what online content and abuse rules will or should look like. But we need civil, safe spaces for free expression *right now*, which means that someone must create and maintain them. Why not you?

[9] For a discussion of changes that may be coming to the First Amendment, see https://lawreviewblog.uchicago.edu/2022/06/06/douek-lakier-first-amendment/ and https://shows.acast.com/arbiters-of-truth/episodes/the-supreme-court-blocks-the-texas-social-media-law

Authors

Karen Maxim is the Policy & Law Enforcement Response Team Lead at Zoom Trust & Safety.

Josh Parecki is the Head of Trust and Safety at Zoom.

Chanel Cornett is a Counsel at Zoom Trust & Safety.

Acknowledgements

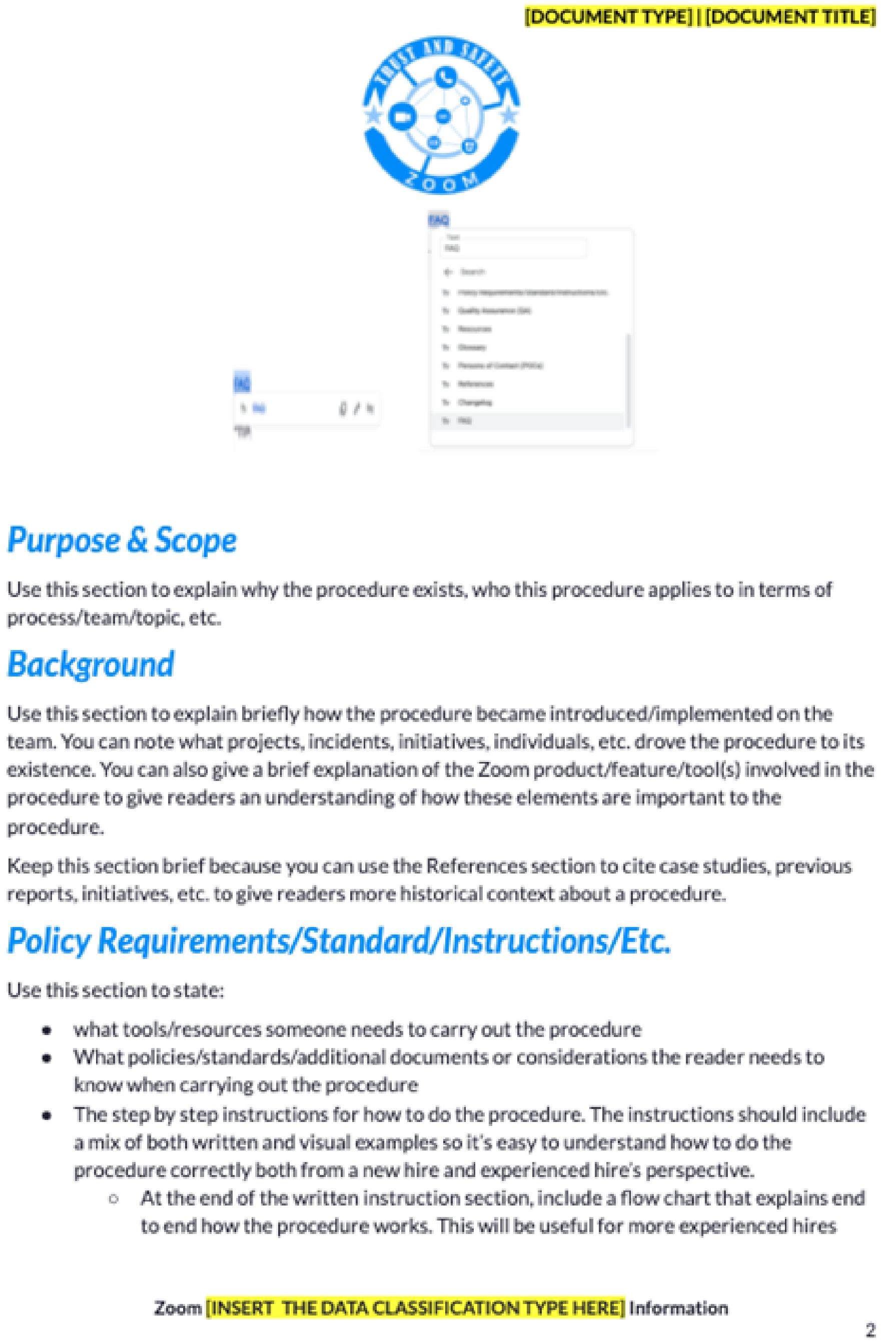
The authors wish to thank Lynn Haaland, Zoom's Chief Compliance, Ethics and Privacy Officer, for her support and guidance during those early hard days and through today. We also want to thank Josh Kallmer of Zoom's Government Relations team for his close partnership and support. Finally, we want to thank Caroline Meinhardt for her editing and feedback.

Keywords

Trust and safety; video

Appendices

A Template for Creating a New Standard Operating Procedure (SOP) (Updated March 2022)


TRUST AND SAFETY
ZOOM
FAQ
Text
FAQ
Search
Policy Requirements/Standard/Instructions/Etc.
Quality Assurance (QA)
Resources
Glossary
Persons of Contact (POCs)
References
Changelog
FAQ
FAQ
FAQ
TIP

[DOCUMENT TYPE] | [DOCUMENT TITLE]

Purpose & Scope

Use this section to explain why the procedure exists, who this procedure applies to in terms of process/team/topic, etc.

Background

Use this section to explain briefly how the procedure became introduced/implemented on the team. You can note what projects, incidents, initiatives, individuals, etc. drove the procedure to its existence. You can also give a brief explanation of the Zoom product/feature/tool(s) involved in the procedure to give readers an understanding of how these elements are important to the procedure.

Keep this section brief because you can use the References section to cite case studies, previous reports, initiatives, etc. to give readers more historical context about a procedure.

Policy Requirements/Standard/Instructions/Etc.

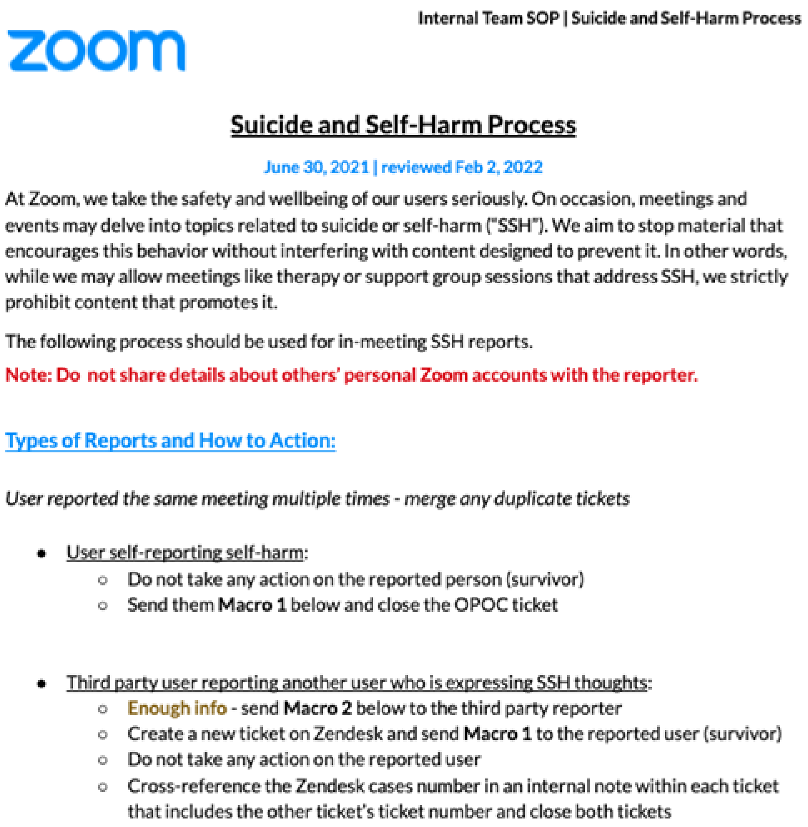
Use this section to state:

- what tools/resources someone needs to carry out the procedure
- What policies/standards/additional documents or considerations the reader needs to know when carrying out the procedure
- The step by step instructions for how to do the procedure. The instructions should include a mix of both written and visual examples so it's easy to understand how to do the procedure correctly both from a new hire and experienced hire's perspective.
    - At the end of the written instruction section, include a flow chart that explains end to end how the procedure works. This will be useful for more experienced hires

Zoom [INSERT THE DATA CLASSIFICATION TYPE HERE] Information

2

B Suicide and Self Harm Process SOP



zoom

Internal Team SOP | Suicide and Self-Harm Process

Suicide and Self-Harm Process

June 30, 2021 | reviewed Feb 2, 2022

At Zoom, we take the safety and wellbeing of our users seriously. On occasion, meetings and events may delve into topics related to suicide or self-harm ("SSH"). We aim to stop material that encourages this behavior without interfering with content designed to prevent it. In other words, while we may allow meetings like therapy or support group sessions that address SSH, we strictly prohibit content that promotes it.

The following process should be used for in-meeting SSH reports.

Note: Do not share details about others' personal Zoom accounts with the reporter.

Types of Reports and How to Action:

User reported the same meeting multiple times - merge any duplicate tickets

- User self-reporting self-harm:
  - Do not take any action on the reported person (survivor)
  - Send them **Macro 1** below and close the OPOC ticket

- Third party user reporting another user who is expressing SSH thoughts:
  - Enough info - send **Macro 2** below to the third party reporter
  - Create a new ticket on Zendesk and send **Macro 1** to the reported user (survivor)
  - Do not take any action on the reported user
  - Cross-reference the Zendesk cases number in an internal note within each ticket that includes the other ticket's ticket number and close both tickets