Google Prodcast Season Three Episode Six

[JAVI BELTRAN, "TELEBOT"]

STEVE: Welcome to season three of The Podcast. Google's podcast about site reliability, engineering, and production software. I'm your host, Steve McGee. This season, we're going to focus on designing and building software in SRE. Our guests come from a variety of roles, both inside and outside of Google. Happy listening. And remember, hope is not a strategy.

[JAVI BELTRAN, "TELEBOT"]

Hey everyone. Welcome back to season three of The Prodcast from Google. This week I'm joined again by our co-host, Jordan.

JORDAN: Hey.

STEVE: I'm Steve McGee, and we have two guests this week. We have Sarah and Vrai. Why don't you guys introduce yourselves?

SARAH: Yeah, absolutely. My name is Sarah Butt. I work with Salesforce with a team called Centralized Incident Response, which is part of Salesforce's reliability engineering organization.

VRAI: And I'm Vrai Stacey. I work in Google's SRE on our internal incident response tooling.

STEVE: Awesome.

JORDAN: Awesome.

STEVE: So in this episode, we're going to focus on incident response and specifically about the tooling and the software that teams might want to use or might realize they don't have and they need to build themselves. But first off, what do we mean by an incident? There can be lots of things. Basically the stuff that goes wrong and everyone has to go, whoa, wait, what was that? Sometimes there's security incidents, and they can be a little bit different from reliability incidents. We're going to focus today on reliability. But Jordan had an interesting one that I think he kind of helped set the stage. Let's hear your favorite one, Jordan.

JORDAN: Yeah. So many years ago at Google before I joined the SRE TPM team, I worked on the

support team. And we got a funny ticket from an SRE that said we needed to get additional Wi-Fi access points installed in the bathrooms, because the old building was not allowing the data to flow through the walls and you maybe missed a page here or there just to meet the SLA. We have to get better Wi-Fi.

STEVE: That counts too.

VRAI: Which may have been me if it was London because we had that exact same problem when we moved into our new building. And you had to find a secondary before you went to use the facilities, which was suboptimal.

JORDAN: Wow.

STEVE: Was that PS, Vrai?

VRAI: PS, yes. So our Kings cross campus.

STEVE: I was there too. I remember that actually. I remember moving into that office and there were some issues. And so this is actually a really great funny example because who saw that one coming? Was there a playbook for this? Probably not. Some complexity has arisen through outside mechanisms that we didn't see happening. So that's awesome.

JORDAN: It goes to show how important the process is when it comes to communicating and talking with your team when having an incident. So why do you need to get paged in the bathroom? What is incident response, and why did you need Wi-Fi to be the person who fixed whatever was going on that day?

VRAI: So at the time, I was on call for some of Google's network infrastructure, predominantly the part that connects our Google's internal systems to the internet. So it's kind of critical for what we would do, how we do business. And, as you mentioned, the building, we just moved into it, had terrible cell phone coverage, terrible Wi-Fi. And we had a three-minute SLO. So we were expected from having been paged to be in front of a computer fixing a problem within three minutes. That meant if you needed to go and spend a penny, you basically needed cover if you couldn't get your coverage. Otherwise, you could be incurring ad serving could be down. YouTube could be down, and you would have no idea until you got back into somewhere where your phone could shout at you again.

STEVE: Yeah, an important point is this implies that you were the one on call. It wasn't like drop a

message into a common thing. It was like a single person gets a single message. Is this common do you think, or do people start out this way or is this something that evolves over time? Maybe you guys can tell us about your picture of how these systems tend to work out in the world and what systems are required to make this work right.

SARAH: I think the single team, single person rotation, that's something that I've definitely encountered, especially I think depending on size of company, or resourcing, or how critical it was. And one of the things that I've tried to encourage as I've led teams or set up rotations is not having a single point of failure in terms of just one person getting paged. Now that might be we have a person that gets paged and something happens and that rolls through them for some reason, and then two minutes later the next person is getting paged, or it might be two people get paged in parallel. And that's expensive to staff. I'll be the first person to say that's expensive to staff. And some of the things that I've had to talk through with people is like, can you partner? If you are an SRE team that partners with service learning teams as well, can you all partner together to share the page? Can you partner with similar teams where even if you don't have the deep expertise, both teams can do troubleshooting on each other's service just to get more people in rotation? Because I think there's a very human cost of rotation that comes into the need to also provide robust coverage. I'm very lucky, I would say, on the team that I'm on right now that we do-- we always have a primary and a deputy, and they're both getting paged for every incident that we're taking, and they're always on and supporting. To tie that back to the tooling, though, you also need tooling that allows you to do that. Whether that's a quick override or you need the ability to have paging in parallel. So I think figuring out the way to do that is equally important because missing a page feels really terrible. I've always when I've been on a call, I've felt very responsible for the systems I was on call for. And there was nothing worse than feeling like I missed a page and let my team down.

STEVE: Totally.

VRAI: I mean, absolutely on that front. I did have a secondary, just to be clear, and that was the person I would have handed off to. But, again, it's not so much the I will have to pull through, and it'll be fine. It is there's a kind of personal pride if you are on call about making sure that you are on call, and you're ready to go, ready to leap into action if and when required. But, yeah, absolutely I agree also on the tooling. You really need tooling that will fit your needs and your rotation not having to build your rotation around suboptimal tooling, which they may not allow parallel paging or proper portals and whatnot.

STEVE: So you get the page. You have this tool that lets you someone take over or make sure that you don't get paged while you're on vacation. So I can see how tooling works there. And that arises through figuring it out. We have a team of people that have certain demands because they're actual humans and they're not just response robots. What comes after that? So once you get the page, you got to do some stuff. And same thing. maybe it's not just maybe you're working with other-- maybe you're actually collaborating. Can you imagine it? Can you tell us a few more things that come up in terms of, what have teams or companies developed in order to assist that process of right after you get paged, then what? Do you just do it in the dark in your terminal and that's it? All you really need is like a Unix system, or is there more to it than that?

SARAH: I think the way that people define incident response tooling is always really interesting to me, because there's several lanes of it, and there is the traditional lane, which is we use whatever chat platform or voice platform paging platform. There's a foundational level there. I think we're seeing in industry a new level of tooling as well that is specific incident response, whether that's thoughts and automation or tools that help in the post-incident review. But we're seeing that as an actual market segment now where vendors are coming into that, and I think internally teams are building that as well, depending on the size of company and the resources they have. But I always encourage people that like your incident response tooling is much broader than you think it is. So it's also your monitoring and your observability and the dashboards that you have access to and how do you interface with your customer support folks. So some of the tooling that I've seen built is things that really surfaces, information that responders need to make those quick triage decisions. So it's like if you have a failover and you need to know if your standby site is ready or your standby is here or whatever it is ready, surfacing that quickly to responders.

If you need to know if you're getting customer cases in certain regions or not surfacing that to responders, being able to look at something, a CUJ or a Critical User Journey, I think all of those are also tools in your system. And as we learned earlier, Wi-Fi is a tool in your system. These system boundaries are much fuzzier than we think they are.

VRAI: There's a particular issue certainly for companies like Google where we have quite high levels of vertical integration. So being on call for say, the network means that in theory, something you could be on call for could fail, and that would remove your ability to use a bunch of your tooling depending on how deep the stack was. So to go back to collaboration, nowadays, working on the tooling I do, we

tend to collaborate using Google Chat, Google's version of Slack effectively, ChatOps being the big thing. Back in the day, you couldn't rely on chats being around because we were on call for something that sat below it. So we would actually use IRC. Do you remember that very basic minimal stack? But the first thing you do is you look at the problem, triage it, and just start pulling in people. You're part of a team. And we always adopted a view that if the on-caller needs your help, you drop what you're doing and help the on-caller. In the days when everyone was in the office, we could generally see by looking at a team how difficult their problems were by the number of SREs crowded around a single monitor. Whether they be a four or eight SRE issue.

JORDAN: Oh wow. It's like a different measure. You have the five nines, but you also have the eight SREs.

SARAH: Then you start setting severity based on the number of SREs. It's a different in the matrix. You've got like your intensity that you need your urgency and your priority and then the number of SREs around the monitor.

JORDAN: Exactly. How many does it take to change that light bulb?

STEVE: I think it might be a log scale.

JORDAN: Yeah, exactly. So I was thinking about incident communication, how you communicate to your teammates. Like, hey, I'm working on this. This is the current update. Has anyone had experience also having to do the final layer of that which is like end-user communication? What do you tell the communications person at your company? For Vrai, it's probably Googlers who need to know, but I'm wondering if either of you have experienced saying like, hey, our service is down. How do these messages get made? And what do you have to consider? And what information is oversharing, under-sharing? Go for it.

VRAI: So in both of my major roles at Google, both being in the traffic on call and working on the tooling, most of my customers have been other Googlers, at least directly. Obviously, there is a transitive dependency, especially in the network side of things, where an outage could affect large chunks of the world if you broke something really, really badly. Though, these days with obviously cloud being a huge part of our business, we do have to be aware that what may appear to be a small internal outage may have knock-on effects that are knocking out something customers have paid for and therefore are going to be legitimately angry if it's not available, and making sure that the data

flows to them in a timely manner at a level that's useful for our customer care support people is becoming more and more important. And the tooling we're having to build more and more of it to keep this streamlined. So trying to move so that people can focus on what they need to be focusing on without having to do the busy work of data transfer and having that handled for them automatically I think is going to be increasingly important as we go forward.

SARAH: It's interesting to hear you say that because that echoes a lot of what I've seen at other companies as well. So when I started doing incident management, I was working on a large e-commerce platform. So that obviously had external users and a lot of whether that was support agents or success managers, account execs who were very interested in what was happening with the incident. As I've come to Salesforce and other companies, I've had similar where there were definitely layers of external communication. And one of the things that I've worked on recently is like, how do we provide sufficient communication and transparency to-- in the case of Salesforce, it's to our teams that equip our customer-facing staff while at the same time keeping what we call the bridge airtime clear for engineering and mitigation efforts. And so, there's a lot that we've tried to do as far as the tooling goes to both reduce that overhead but also create parallel channels and paths of communication. So can we have the voice bridge focused on the engineering mitigation, because that's where we need the most high bandwidth communication to happen at that time? And then, can we have a separate Slack channel that's talking to our customer support folks-- or in our case, we call it the CIC-- to get them that information? And then, in some cases, if we need to have that, how do we flip them? But basically, how do we create this adaptive system with multiple channels of communication that allows people to focus?

STEVE: That's really interesting. It's something that I've seen that I feel like I'm hearing as subtext in what you're describing is like there's the incident itself. Like fix the broken blah, and then you've got a bunch of them over time. But then you have to go back and be like, every time we try to fix the blah, we struggle because we have this meta-problem. And either it's the tooling or the lack of tooling or the we didn't even realize we needed tooling. But basically, it's some systematic thing that is imperfect in some way or maybe the way we're adapting because our team got bigger or we supported more services in the last six months or whatever. So this I've seen come up when doing like retrospectives and especially if you're doing more than just one incident but like doing a retrospective across many incidents at once. Have you guys seen teams do this analysis a lot? Does it just happen? Is it just kind of like people just make a guess as to what we should build or buy next when it comes to tooling? Or is it more explicit like let's do a meta-analysis of the last months, months, months of incidents across

different teams and figure out what's missing? Like how deep do teams go when they're trying to figure out how to make this system of response better?

VRAI: It becomes tricky when you ask people to try and fix both the process and the systems because they will tend to over-index on their own particular use cases, which is fine if you're the only team in the company running on an entirely custom stack. But it becomes much trickier when you're trying to build something that works for a team dealing with the lowest level of a network stack up to someone dealing with this very broad customer-facing product. And I absolutely agree, your retrospectives or postmortems, as we call them at Google, are a critical factor for this, as is the kind of meta-retrospective where you take the outcome of many postmortems and try and find those common factors that you can identify and then streamline. And that's really how we do a bunch of our roadmap planning for our internal tooling. What can we fix or improve that will have the biggest impact, either in terms of the most people or just, are there areas that are causing customers, the company in general, just huge amounts of pain that we need to get on top of?

JORDAN: I'm curious, do you ever end up in situations where groups have opposing needs based on their incident response styles? And how do you deal with the trade-offs in that discussion?

VRAI: Yes, all the time, both from legitimate needs and perceived needs. Firstly, I think you have to accept you cannot solve everyone's problems optimally, even if you had infinite time and infinite resourcing. It's short of building completely discrete products for people. You would just end up with a spaghetti mess that no one could understand and no one could use optimally. So for us, we're focusing on the majority issue. We're focusing on the 80%. If we can get 80% of on-callers and people doing reliability reviews and people doing customer care in a happy place with the core product, that's fantastic. But the remaining 20% it's extensibility and customization and making sure that people with legitimate business cases that are different from the norm, that we're able to support them. But at that point, we're kind of saying, we're delegating this functionality or this part of the system to an extension that you're going to run. You're going to look after, and that should be perfect for you, hopefully. I really don't see another way around it because in any large company, you're going to have such opposing viewpoints and opposing needs that one monolithic product isn't going to do the job.

SARAH: That echoes an experience that I've had at a previous employer where I was dealing with a vendor, and we had a very specific use case. And I knew that they weren't going to be able to build

exactly for us because it didn't make sense for the rest of their customers. And I remember talking to them and just saying like, all I care about is giving me the best APIs possible. If you will give me an API, I will have my engineers build the other piece of it. And I called it a bridge at the time. I was like, I just need a bridge. Like just give me the pieces so that we can make the bridge halfway, and we'll take it. And I do think that when you get into some of the build versus buy considerations, people kind of assume it's this binary. We're either going to build or we're going to buy. And there are often cases where you're going to do both, particularly in these larger and more complex implementations. And so, getting people to the point where they're willing to wrap their mind around that we're going to buy a foundation and we're going to build our needs on top of that versus just assuming it's going to come for us perfectly out of the box.

STEVE: Yeah, I often hear companies ask like, what systems should we use for incident response? And my response to that is like, I don't know. There's a lot. There's so many.

JORDAN: Whatever works, really. Whatever you have.

STEVE: Yeah, but it reminds me a bit of the Unix utility model of just like I need sed and cat and grep. They all work together. There's not just one-- I don't just type "Unix" and hit enter. There's a bunch of things that I need at different times. And so, it's not clear that you're going to have one tool to rule them all. And also, the tool that you need, you might not even know that you need it yet until the hard day.

SARAH: There's one flip side of that I wanted to add because I think it's easy when you're in a smaller company and maybe you don't have a budget, or you're in a company that's very conscious about security things and incidents, and they don't want to touch tooling. I've talked to people who get very discouraged who are like, I hear about all these great tools, or other people have the resources to build these tools, and we just don't have that. And one of the things I would actually say to those people is like, you have tools. You're just not calling them tools. I have been amazed at big companies with amazing incident response who have done it with Google Docs or with a Google Sheet or with a Slack workflow. And so I just think just in case there's anyone listening who's like, this doesn't apply to me because I can't get budget approval or I don't have a team that can build it, you will be amazed at what you can do with a Google Doc because it's about how the tool fits into your process. It's not just if you can't buy the tool, you can't have good incident response. So maybe that's a little bit of a hot take, but just in case anyone's in that position, because I've been there before and been bummed out, and it really helped me to recognize non-traditional tools or foundational tools that like a spreadsheet

is a wonderful thing.

JORDAN: If nothing else, SREs are very scrappy. So while we have the flashy tools that are like, here's the incident response management tool that we use to show the severity, how many SREs it is, what the impact is, et cetera, where is the actual work getting done? Probably in the group chat, probably in IRC, probably in that Google Doc and then cut and pasted into the flashy thing. So I completely agree with you about using what you have and building the right process to synthesize all that information properly.

SARAH: I am never going to tell people that as someone who has held various roles in terms of owning budget for buying tools or having resources that could develop tools, like I'm never going to tell you that some of those things didn't make my life easier. They absolutely did, particularly in these larger, complex environments with more incidents. But I just fundamentally also push back on anyone who says like, well, we can't improve things because we can't buy a tool. That's the piece of it. Did they help? Absolutely. They're lovely to have. But you don't have to have them to have a great incident management process.

VRAI: Well, process is the keyword. Tools can assist you in having a good process, and they can make it easier. They can remove a lot of the manual work, but tools can't give you a working process. You can have the best tooling in the world, if you're not using them properly, your incident response is not going to be good.

SARAH: Exactly.

VRAI: So just to back up a bit when Steve mentioned the Unix approach, I would put some caution towards that, though, because that's effectively what Google used for many, many years, which was a series of components that have been bolted together over time. If you understood them, they were great. The problem was, it was taking months to train people to be able to use our on-call response tooling. And I think the trick is when whether you're buying or building tooling or as Sarah pointed out, using a combination of both by extensibility, you need something that is powerful enough to support your process but not so obtuse that only half your people can actually use the thing properly because then you get frustration. You get accidental misuse of it. And then, the tool becomes a detriment to the incident response process.

STEVE: Speaking of amazing tools, it wouldn't be a podcast if we didn't mention AI, but we're only going to mention it briefly. Have we solved incident response with AI yet? Vrai, go. Like, easy answer. Nailed it.

VRAI: Obviously, no. Otherwise, I wouldn't be working in incident response tooling. AI is a tool like anything else, and it's a tool that's making radical and great progress. And I think it can help remove a bunch of the toil from being on call, capturing, summarizing, categorization. It's great at that. Can it also help with very simple automatic rollbacks? Sure. But it's not creative, at least not at the moment. And until we get to that point, are you going to trust your company's crown jewels, the thing that makes all your money, to a system that could just make things far worse without human oversight? I believe we're on a road where sure, we'll get to a point where AI may become the predominant mechanism by which you deal with this, but I think we're some distance away from that without a radical leap forward.

STEVE: One thing for the humans in the room and not the AIs is for a hot take moment. I'm wondering if each of you could define the term SEV for me very clearly and with no ambiguity. And the best part about this is [INAUDIBLE] is looking at me funny because Google doesn't use the phrase SEV even though the entire rest of the industry does. So like, what does it mean for an incident to be high priority? Like, what's the difference between P1 and P4, SEV1, SEV2? This is obviously a trolling question. But why is this a hard question? That's my real question.

SARAH: Can I first refer out to there was a really fantastic presentation at Esri Americas this past year on what is incident severity but a lie agreed upon. It's freely available on YouTube, so I would highly suggest everyone go watch that. But the way that I tend to define severity is they serve your company as an indicator of an idea of the impact and the complexity involved in an incident. And so every company has different-- I've been at companies where we ran SEV1s every week. I've been with companies where we saw SEV1 once a quarter, and I never saw SEV2 or I saw one during multiple years of the company. Companies get to set them. They mean certain things within the company. But I do think it's interesting how I think there's a comfort that we try to get from severity level in that humans have this need to categorize things very neatly. And if I can put this neatly in this box, I'm dealing with this novel situation, and it's overwhelming, or I'm concerned about it, but I can put it neatly in this box. There are often a lot more fuzzy. And so, what I tend to tell people is like you call it the pink pony incident severity for all I care. All I need to know is, do you need-- in this case, CRR and my team, do you need our support for the incident? Then great. We will be there to support the

incident. And we will be there to provide an incident commander and whatever piece we need to do. But I think severity is very much an organizational construct, and it's a model. All models are flawed, but some models are useful. That's like a bad appropriation of that quote. But I think that's what I would point to as severity, is like they are useful models in some cases. But there's a lot more there. It's a great SRE talk though. I would encourage. Em did a fantastic job with it.

VRAI: I would agree that severity, it depends. Because it depends on what you're measuring. But ultimately, it is a communications tool. It's a way of saying like if it's a huge outage, to use our internal terminology, then that means all hands are going to be on deck. We know it's having an impact across a wide area, whereas if it's a negligible outage, it probably means only on-call is going to care. And if anyone is glancing, they won't worry about it. But there's no one size fits all, because there's no one size for any organization, any PA, any team.

STEVE: Yeah, I've heard a good definition of it as Sarah said, you can call it pink pony, whatever you want. But as long as people understand what the outcome of each of these levels means, in some companies being at P1 or SEV1 means you have purchase authority or means you're allowed to bring in other teams, whereas if it's an SEV3, you're not allowed to page anybody else or something like that, like whether that's a good rule or not, whatever. But the point is, what is the meaning of assigning these letters and numbers like to the rest of the org or to your own teams or what you're able to do or not do or whatever? As long as that's well understood within a team, I think that's important. And then the next phase of that is to make sure you know that you're allowed to change it. Like once something is a SEV2, you can change your mind. You can be like, actually, no. It was a two. We're good. I joke with teams that I consult with, like when was the last time you demoted an incident? And generally, it doesn't happen. Like if it gets to SEV1, it's that for life, which is a bummer because it can change. You can understand it better.

SARAH: I also try when I'm coaching people on incident response to encourage them not to get overly caught up on that during the incident. Like if you're spending more than a few minutes doing that when you should be mitigating-- I shouldn't say should. That's probably too strong a word-- but it's not worth the conversation when you're dealing with the impact. So I tend to say severity serves the incident. If the severity is serving the needs, we have the people. Even if we have to retroactively say actually we're going to upgrade that or we're going to change that, I'm fine with that if it means that it kept our airtime open to work towards the path to green. So it's just one of those things I tell people like it serves it. There's cases where I've said, OK, we're going to go SEV1 with this, and it's not going to

be a discussion right now. And if it needs to be a discussion in the retro, that's fine. But what I need right now is the mechanisms that SEV1 opens to me, whether that's a path to additional teams, a path to legal, expedited, whatever it needs. And we can talk about later, do we need paths for those at lower severity so that I don't have to push the big red button to get that? That's fine, but we're not going to do it during the incident.

JORDAN: Yeah, definitely.

SARAH: There's another SRE talk that's fantastic, and I think I would be remiss to talk about this and not mention this talk to people as a resource. It's on the future of above the line tooling. It was given a few years ago. It's available on YouTube by John Alspaugh and Richard Cook. And they have incredible insight into some of this. One of the big takeaways that I've continued to use is this concept of clumsy automation, which is automation that increases workloads at a high workload moment for the responder and decreases it at a low workload time. So the academic paper that this came from talked about people using this in aviation. And if you had to do a bunch of stuff during takeoff and landing that made cruising easier, it actually wasn't useful automation because takeoff and landing are these high cognitive workload times. And so I would just say that talk is very worth listening to, particularly if you're going to be designing tooling or even having influence in your company's tooling because it gives a lot of insight from not just the tech industry, but also other industries that have dealt with tooling in these high acuity, high cognitive workload situations.

JORDAN: OK.

VRAI: I think there is a lot aviation can teach incident response in general just because it's an area of industry that's had so much time and money invested in making it safe as physically possible. There's a huge amount the tech industry can learn from how they deal and how they respond and their human factors training by how you can make sure that everyone involved in an incident feels that they're allowed to speak up and that their input will be taken on board, so you don't end up with just one person's tunnel vision leading you down the wrong path. I do have a bugbear about learnings and whatnot, but I could go on for far more than this podcast runtime about it.

STEVE: Do you want to do a quick dive into that? So specifically, you're talking about retrospectives and getting learnings out of a system.

VRAI: I mean, predominantly I find that a lot of the incident response tools and a lot of the response

process discussion concentrates just on the triage and mitigation stage. It's about how serious is it, get it fixed. Stop the bleeding, to use the SRE terminology that we have. But if you're constantly solving the same problem and the same outage, yes, you may have a very slick process. By the end of it, you may have automation to help you. But fundamentally, an outage that you don't learn from is a failure. An outage that you learn from and that you take steps and you're able to prevent an identical and similar outages occurring in the future, it may have been costly, but there is a positive side to it. And I think we need to re-balance somewhat the investments in general in incident response into let's not have the same incident happen twice. There will always be new and interesting ways to break systems as long as people are using them and people are writing them. The black swan event, so to speak, would always be there to destroy the best-laid plans. But an investment in making sure that you're constantly improving your process, constantly improving your systems and just never falling down the same hole twice is worth vastly more, in my opinion, than just piling it all into having the slickest possible mitigation system you can possibly get.

SARAH: I think one thing-- because I have definitely consulted with companies where they said, we don't have time for that. Everything is on fire, so we don't have time to do that. And the rebuttal that I've made-- and I need to credit John also with this, but he would always tell me, an incident is like the unplanned investment that you've already made. Like you've already made the investment. So I would tell you, what does it cost for an engineering hour? How many engineers did you have involved? Like you have already invested that. It has already impacted your customers. If you've already made the investment, it's worth it to actually get insight from that. And that's a skill set that takes some learning to do that and do that in a way that's psychologically safe and do that in a way that's valuable and generalizable. But I really would encourage companies that it's a worthwhile investment. One of the most encouraging things that I've seen is how I feel like as an industry, we're moving away from MTTR right now as the single be-all and end-all metric, and there's a lot more we could talk about that. And I will, again, just say like Courtney Nash did amazing work in the void report, and it's worth reviewing that. But seeing how processes and tools can get us insights beyond that in a way that is still low overhead is really an exciting next chapter in incident processes and tooling for me.

JORDAN: So, wow. It sounds like it's a little bit fun but also terrifying to be in the pink pony club of critical issues affecting a service. So let's spend just a few minutes-- if you don't mind-- telling us about some of the most really intense incidents of your career and any sort of takeaways that you might want to share.

VRAI: So in terms of probably the incident that took the most time to put everything back together, again, and I'm going to have to fuss over some details here because I can't remember off the top of my head how much we've announced about it. But this happened back when I was on call for the networking side of things. I came into the office one day, not on call, to see our on-caller there. And it turns out a large portion of our serving stack had disappeared overnight. And by disappeared, it seems that the automation that was there to securely remove data when we turned systems down had gotten a bit overenthusiastic and just gone through and very, very fast, purged them. All the machines were still there. They were still connected to the network, but nothing was running on them. And it's one of those areas where, as mentioned earlier, automation sometimes does more harm than good. In this case, there was a tiny one-line bug in the automation. That meant if you told it, gave it a list of machines to clean up, it would do it. If you gave it an empty list, it would be like, "An empty list you say? Well, that means I'll just destroy everything." And it had been through and just removed this huge chunk of the fleet, which is kind of a testament to how, back in those days, overcapacity we were, that really I think only those people who go around constantly pinging us to check latency actually noticed there was an issue. But it so comprehensively broke a bunch of stuff that we had to manually reinstall huge chunks of the fleet to get it up and running, which took, I think, an entire team. Sorry, I forget how many SREs there were to count the problem. But it just struck me as one of those things that we built this tooling to save time, and now, due to one tiny error and the lack of having a human provide oversight, it had caused far more trouble than it had ever prevented. But this is why you have learnings. This is why you have postmortems and retrospectives. Don't have a system whose default behavior when you pass it an empty list is to just go on the rampage through your infrastructure. Seems obvious now, but apparently, it wasn't when the tooling was being written.

JORDAN: Yes.

STEVE: Gradual change is important. That story is actually in the SRE workbook. It's in a chapter. It's a case study in a chapter.

VRAI: OK, so it has been mentioned.

STEVE: You are certainly allowed to talk about it. It's in the book, or it's in one of the books. How about you, Sarah?

SARAH: I don't know that I can give you a specific example, probably because my employer prefers me to adhere to my NDA and past employers as well. What I will say is, I've been doing this. This has been

the predominant focus of my career for over a decade. It has been incident management, incident response. So I've seen a lot of incidents. I've also trained a lot of people on this. And when I talk to them, I guess my takeaway from these really bad incidents, the ones where you're like, we might end up on the news for that incident, like that was real, is always to tell them three things. And I normally tell them proactively as I'm training incident commanders or as I'm mentoring people, like you need to have these three things in place when you prepare to be on a major ICO rotation or whatever your company calls it. You need to know the people or the person in your company-- and they might be in your team, they might be a mentor outside your team-- that you feel like you can talk to to unwind from an incident. They are under an NDA. You can say anything about the incident. That person or people is so important. I really encourage people to build connections in the industry, whether that's on Slack or your social network of choice or at conferences, but developing friendships with people who work in this space but are not with your company is so key for that outside perspective. And I think I have finished up some just really adrenaline-flowing-- because the thing that I tell people is like, you can seem really calm as an incident commander. When you're in these really bad incidents and they're incredibly novel, everybody's a little bit terrified. If you're not terrified, I would like to know more about the ice that runs through your blood, but it does not run through mine. Having that person that's outside your company and outside perspective that you can go to and say like, I can't give you the details, but I just ran a heck of an incident, and I'm trying to unwind, and that kind of bouncing off of the person that's in the industry and gets it is so critical. And my spouse would probably say that he deeply appreciates them because otherwise, he would hear a lot about incidents over the dinner table. The last thing I tell people is to find something separate that gives you perspective because you deal with these things, and they're very real, and they impact a lot of people, but it's easy to make them your whole world. And so for me, I go volunteer at a children's hospital. My dog is a therapy dog. It gives me so much perspective every week. And that's not to say that that's the reason I do it, but it's one of the things like the secondary effects that I'm very thankful for. It's like finding a way to say everything about this is serious and it can be all-encompassing, and I want to put my whole heart and self into it, but I also want to have that perspective that, yes, the sky is falling, but probably only for one particular cloud.

STEVE: Let me throw one last curveball in there for my story. Instead of an incident, I want to talk about how I've used incident response in my life, which is I use it to help plan family vacations because they are like an incident in the making, not necessarily in a bad way, but like there's a lot of stuff going on.

There's a lot of people collaborating. There's outside events. And just having a way to think through it and collaborate effectively with your peers. We have each other. My wife and I are like secondary and primary on call, and we switch right throughout the week's vacation who's making sure we have train tickets and things like that. We have the collaborative way to find those assets when we need them, like who has the app with the right check-in code on it and things like that. And it sounds silly, but I honestly think that being able to handle this type of slightly pressured situation with a group of people transfers really well into the world.

JORDAN: Deeply. Yes.

STEVE: We don't have an incident management system in terms of tooling, but the process is there and the way of thinking. And the not panicking and the not blaming each other and the not running away screaming, because you know it's got to get done. We got to get on the train. We got to get the service back up. It's the attitude.

VRAI: Do you publish postmortems at the end of your holidays?

SARAH: Steve, I have to ask, because you asked us earlier. So what's the difference between a SEV1 and a SEV4 on a family vacation? Do your vacations have severity levels?

STEVE: So there was one SEV1, which was when the train tickets did not exist. And actually, I was a solitary responder in this event. It was not a family vacation. I was just like on my own somewhere. The train ticket did not exist in my backpack where it was supposed to be, and the mitigation was to go buy another ticket immediately. And it turned out the train ticket was there. I just didn't find it. So then I had two train tickets. It was an over-and-above response.

SARAH: Redundancy.

STEVE: Yeah, it was great.

JORDAN: Did you run your postmortem because you used resources?

STEVE: The thing where I got lucky on that incident, yeah, I did actually write a postmortem for this just because it was funny. The place where I got lucky is while I was standing in line to buy that redundant ticket, I stood in line behind the only French astronaut that I know of. And we talked for a while about how he fixed the Hubble telescope. It was pretty rad. He was carrying a bag that had like [INAUDIBLE] on it and things. So that was fun.

VRAI: That's got to be worth losing a ticket over.

STEVE: I think so.

JORDAN: I think it was on purpose, actually.

STEVE: That's right. So you never know. Well, thank you both so much. I think incident response is one of the things that people think about maybe first when they think about SRE. But the depth at which you can go into it and the amount of detail is not always obvious. So I'm glad we got a chance to go through that today. Anything that you want to add, like places to find you on the internet where people can follow you or learn your favorite joke or whatever?

SARAH: I am not cool enough for most social media, is what I tend to tell people. But I am on LinkedIn. And I'm always happy to chat. Like I said, I've spent most of my career doing this. I deeply love the incident space, so I always tell people, feel free to send me a message on LinkedIn. It might take me a bit to reply depending on how life is going, but I do try and reply. I always love chatting and getting to know people in the incident space. Similarly, if you see me at a conference, I'm always happy to grab lunch together and chat. So that's how you find me. I'm not cool enough for most of the social stuff anymore.

VRAI: I am also not on most social media stuff, but I am on LinkedIn, to the best of my knowledge, the only Vrai on there. So I'm not too hard to track down.

STEVE: Nice.

JORDAN: Nice.

STEVE: Cool. Well, thank you both. This was a blast.

SARAH: Thank you. This was great.

VRAI: Thank you.

JORDAN: Thank you.

[JAVI BELTRAN, "TELEBOT"]

JORDAN: You've been listening to Prodcast, Google's podcast on site reliability engineering.

Visit us on the web at sre.google, where you can find papers, workshops, videos, and more about SRE. This season's host is Steve McGee with contributions from Jordan Greenberg and Florian Rathgeber. The podcast is produced by Paul Guglielmino, Sunny Hsiao, and Salim Virji.

The podcast theme is “Telebot” by Javi Beltran.

Special thanks to MP English and Jenn Petoff.