**Background:**

A fast-growing company called WorkSafe helps large corporations monitor their shared Google Workspaces. Corporations have the legal responsibility to act on any evidence of workplace harassment. With Workspaces, they might have evidence of harassment on their servers without “knowing” it, in the traditional sense. But their lawyers say that when employees harass using chats or shared docs etc., the company may be be found liable in court for scanning that stored data for red flags. WorkSafe does the job: their slogan is “Make your Workspace WorkSafe.”

The university reached out to WorkSafe to monitor their Google Workspaces. They had a similar concern, that they might be held liable if they possessed evidence of any harassment but did nothing. So they propose a custom-built algorithm which they will rebrand as “TigerSafe.”

There are some things that make this different than WorkSafe’s previous jobs, though. First, in order to comply with FERPA (the Federal Educational Rights and Privacy Act) the university will have to keep WorkSafe’s algorithm and all the data it generates on their own secure servers. Second, since courts have recently said that universities have special responsibility for student mental health, they also want the algorithm to flag concerns about that.

Some engineers at WorkSafe were a little unsure: their language-detection models were trained via supervised learning to detect harassing language in particular, not other problematic communications. But their board of directors wanted to expand into the university space, seeing an opportunity for hundreds more clients. TigerSafe was built and deployed.

**Sprint One:**

Within a week, TigerSafe has hit a snag. University servers are run by their IT people, who never talked with WorkSafe to make sure that system updates would be compatible with TigerSafe. The servers updated to a new OS, and TigerSafe immediately crashed. The University needs a solution.

The senior WorkSafe engineers say they could build a module to automatically update their algorithm in response to server updates. That would take time, and cost the university more, but it would be stable and secure. The junior engineers say that’s too old-school. They say there’s an open-source server module that would work, and can be deployed almost for free in two days. Info on the repository where the server module is hosted shows that it hasn’t been updated in a few months, and some bugs have not been addressed. But since it’s an easy and low-cost fix, they can always fall back on the old-school methods if it doesn’t work.

**Should WorkSafe add the open-source module to make server updates, or build its own custom module?**

**Sprint Two:** The WorkSafe board makes the call, choosing the fast and cheap fix. But then the university makes a new request. Students say that most online bullying is in memes. But TigerSafe was only trained on text; it can’t see the words on a pic or gif. Even if it could, it might not catch the meaning, since memes are constantly evolving.

The young WorkSafe engineers propose a fix! On the same open-source repository, they find a module that checks images against a database that tracks online harassers, who often communicate using memes. It could be used with TigerSafe to check for images commonly used to bully, and flag these for university review. And since this module is from the same repository, they know it will work with the server module.

The senior engineers are worried. When corporations use WorkSafe, they limit the emojis or images employees can use. So WorkSafe has never done image scanning before.

**Should WorkSafe use the open-source image recognition module, or tell the university to limit images students can use?**

**Sprint Three**:

The WorkSafe team decides to give the university image scanning. Limiting speech doesn’t seem right, but they feel an ethical obligation to stop harassment in all its forms (memes, words or otherwise).

Four months later, Congress passes an amendment to the 2011 [Protecting Children from Internet Pornographers Act](https://en.wikipedia.org/wiki/Protecting_Children_from_Internet_Pornographers_Act_of_2011), applying it to all web-based services (including those offered by the university). The Act requires internet providers to maintain all data for one year, so that child pornographers can be traced and prosecuted. The new amendment requires providers to check all images against a database of things tagged as CSAM (Child Sexual Abuse Materials) – tags used by law enforcement to track and target anyone who traffics in these. These databases can catalog any illegal materials, from illegal obscenity (revenge porn, or other images deemed unacceptable in some locales) to copyright infringements and so on.

The university feels ethically and legally bound to cross-check everything against CSAM and other databases – if only to reassure themselves that none of their students are involved. They ask WorkSafe to expand the image-check function to do this job.

**Should WorkSafe refuse to expand its product further, or re-code the image recognition module to scan for illegal materials?**

**Sprint Four:**

WorkSafe feels ethically bound to stop CSAM, so they build image scanning. A year later, a few students have been charged with harassment – but the monitoring has mostly reduced bad behavior. And no student has been caught with CSAM.

But now a hacker group known as the Cambridge Collective breaks into university servers. Their m.o. is embedding malware into open-source modules, allowing them access to any server that uses them. The server module WorkSafe chose had been compromised.

The Collective sends a message: “We have your data. Pay us $5 million in 72 hours, or the world will see every spicy pic your students have ever shared! Kiss those Goldman Sachs offers goodbye!”

The university thinks they should just pay the data ransom. But lawyers recommend no negotiations: paying just incentivizes more hackers. Plus, some of the engineers at WorkSafe say that they stand a 50/50 chance of recovering the data anyway.

**You have to advise the university what to do. Do they take a chance on defeating the hackers, or pay them to make this all go away quickly?**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

END OF ROLE-PLAY HERE – If you include a "Sprint 5" here, it should include the results of the team either paying or not paying the hackers -- and then ask the students to reflect on that.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Relevant Sources:

**Malicious plugins in Wordpress**

<https://www.securityweek.com/malicious-plugins-found-25000-wordpress-websites-study>

**Gaggle, a company that scans students’ work for dangerous content:**

<https://podcasts.apple.com/us/podcast/who-watches-ai-watching-students/id1523584878?i=1000568070085>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_