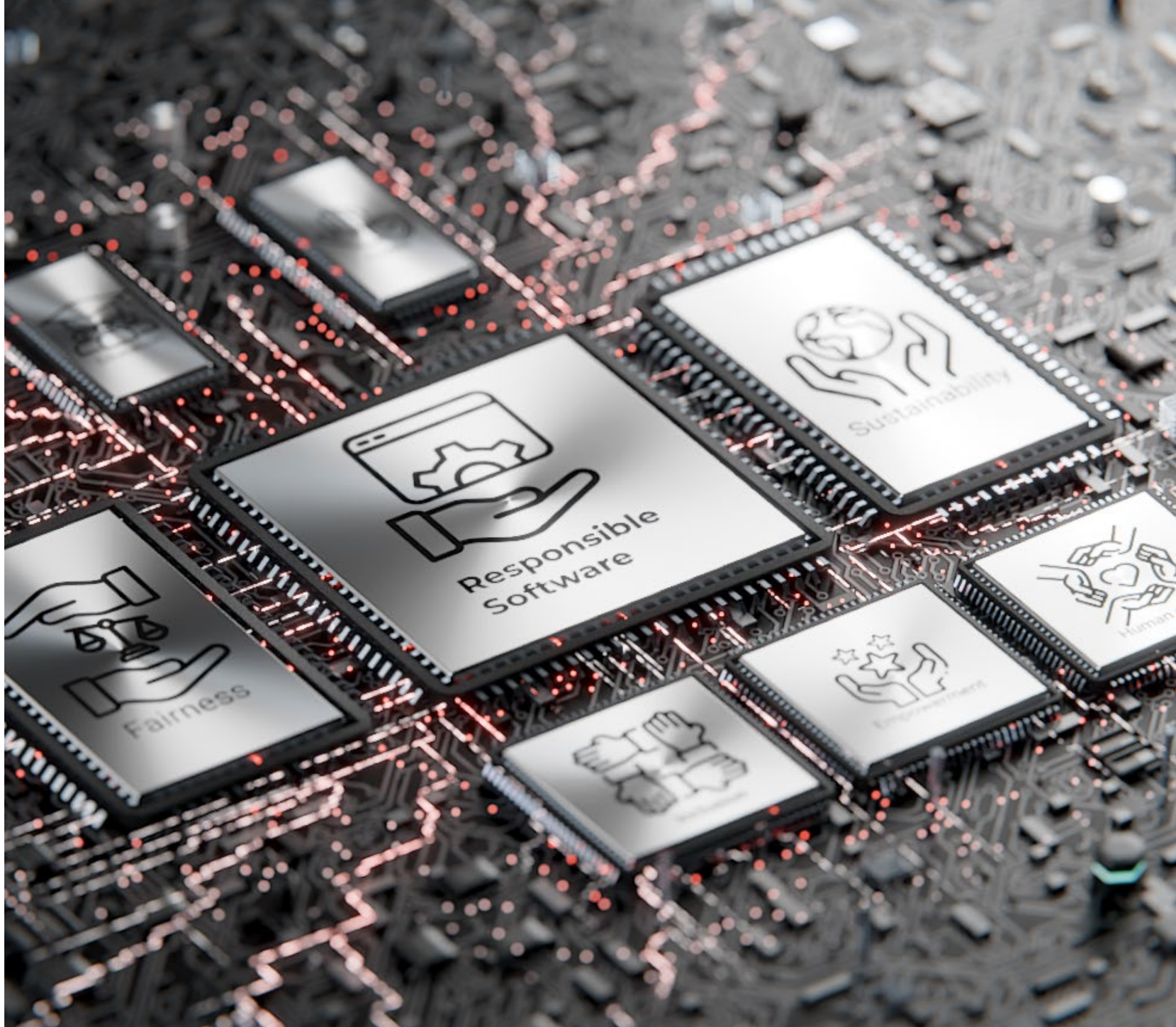
EPFL

Safety 1 Review & Case Studies

22 sept.

Cécile Hardebolle

Responsible Software



Agenda for today

1. Interactive review questions on Safety 1

2. Case studies:
   a) Bad actors
   b) STRIDE
   c) Harm modeling

Autonomous car software - 1

The software of an autonomous car has a 10% error rate in recognizing traffic signs correctly.
We are in the presence of (select all that apply):

- ❌ 0% a. A safety threat
- ❌ 0% b. A security threat
- ✅ 0% c. A safety hazard
- ❌ 0% d. A security hazard

URL: ttpoll.eu
Session ID: cs290

Autonomous car software - 2

Stickers placed on a stop sign lead the software of an autonomous car to misclassify it as a speed limit sign.
We are in the presence of (select all that apply):

- ❌ 0% a. A safety threat
- ✅ 0% b. A security threat
- ✅ 0% c. A safety hazard
- ❌ 0% d. A security hazard

The stickers affect the system negatively
(e.g. may have been placed by bad actors)

URL: ttpoll.eu
Session ID: cs290

Worldwide “CrowdStrike” outage in 2024

This event is an example of:

- ✅ 0% a. Malfunction
- ❌ 0% b. Misuse, abuse
- ❌ 0% c. Unintended use
- ❌ 0% d. Intended use

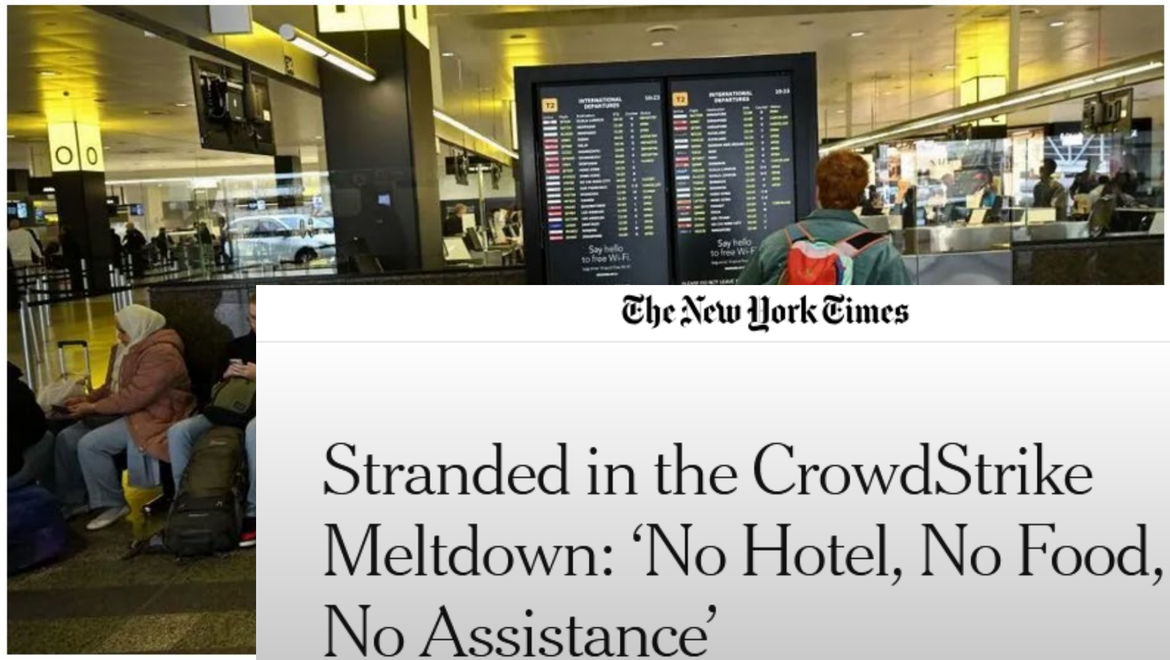
URL: ttpoll.eu
Session ID: cs290

CrowdStrike IT outage affected 8.5 million Windows devices, Microsoft says

20 July 2024

Share Save

Joe Tidy
Cyber correspondent, BBC News



The New York Times

Stranded in the CrowdStrike Meltdown: ‘No Hotel, No Food, No Assistance’

Airlines pledged assistance, refunds and reimbursements to passengers whose travel had been disrupted by this summer’s software outage. Instead, passengers told us, they were on their own.

Bad actors, safety and security

❌ 0% a. Bad actors generate safety issues only

❌ 0% b. Bad actors generate security issues only

✅ 0% c. Bad actors generate both security and safety issues

URL: ttpoll.eu
Session ID: cs290

Bad actors and the 4 scenarios

Bad actors can be involved in (select all that apply):

- 0% a. Malfunction — Yes, if we consider that a bad actor can lead a software to malfunction
- ✅ 0% b. Misuse, abuse
- ✅ 0% c. Unintended use
- ✅ 0% d. Intended use

URL: ttpoll.eu
Session ID: cs290

The “confusing” matrix - 1

We use software to detect fissures in concrete walls before they become visible to the naked eye.
A positive result means presence of fissure.

URL: ttpoll.eu
Session ID: cs290

Select all the correct statements:

- ✅ 0% a. TN = actual absence of fissure, correct prediction
- ❌ 0% b. TP = actual absence of fissure, correct prediction
- ✅ 0% c. FN = actual presence of fissure, incorrect prediction
- ❌ 0% d. FP = actual presence of fissure, incorrect prediction

The “confusing” matrix - 2

We use software to detect fissures in concrete walls before they become visible to the naked eye.
A positive result means presence of fissure.

URL: ttpoll.eu
Session ID: cs290

From a safety perspective, the indicator we should pay most attention to is:

❌ 0% a. TN

0% b. TP

✅ 0% c. FN

❌ 0% d. FP

TP can also be considered as an important indication for safety as it indicates that the software detects properly the fissures

The “confusing” matrix - 3

We use software to detect fissures in concrete walls before they become visible to the naked eye.
A positive result means presence of fissure.
Here is the confusion matrix you get 👉

		Predicted: Fissure	Predicted: No Fissure
Actual	Fissure	60	15
Actual	No Fissure	20	100

<u>What is the False Negative Rate (FNR)?</u>

- ❌ 0% a. 13%
- ❌ 0% b. 17%
- ✅ 0% c. 20%
- ❌ 0% d. 25%

FNR = FN / Actual P
= FN / (TP + FN)
= 15 / 75
= 20%

Case studies

Where to find the cases?

1. Go to **courseware**
2. Find **the case studies** for today
3. Download:
   - The **instruction sheet**
   - The **3 cheatsheets**

Bad actors

Instructions

Remember the notebook examples: “Catter” & Twitter

Individually, identify the potential bad actors for social media in general:

1. Use the motivation categories to brainstorm a range of **harmful actions** that could be performed by **bad actors**
2. Identify the **impacts** for users and for the platform

Share with your neighbor:

- Did you identify the same bad actors?
- Can you agree on a final list?

Which among these are bad actors in Catter?

Select all that apply:

0% a. Sassy posts a fake picture of Dogs invading Purrville

0% b. KitKat re-posts covert Dog propaganda

0% c. Felix promotes cat-biscuits that his cousin cooks

0% d. Tuna posts a series of angry replies to Catnip's post

Bad actor = intention + harm

- Sassy: harm, intention to be determined
- Kitkat: harm, intention to be determined
- Felix: looks harmless but a potential for harm exists (e.g. health), looks like good intention (should be transparent if driven by financial gain, e.g. influencers)
- Tuna: harm (potential harassment), intention to be determined (lack of context)

URL: ttpoll.eu
Session ID: cs290

Overall debriefing of the strategy

- **Motivation categories overlap** and it is not always clear how to classify some types of actions
- The goal is to identify a **range of possible scenarios** that could create **threats** for your system (security) and **hazards** for your users (safety)

👉 **Use as a help for brainstorming**

⚠️ **For the exam:** you will be asked to identify the bad actors that are **specific** to a given app/case i.e. you should make clear the **link** between the app and the actor (not describe bad actors in general)

STRIDE

Instructions

Individually:

1. Match each proposition to a threat category from STRIDE
2. Describe a countermeasure to prevent / mitigate the issue

Share with your neighbor:

- Compare your matching
- Discuss your countermeasures

Debriefing

1.
    1. Information Disclosure (I)
2.
    2. Tampering (T) + Elevation of privilege
3.
    3. Spoofing (S)
4.
    4. Denial of Service (D)
5.
    5. Repudiation (R)

Why do we do this?

- Goal = identifying how bad actors can generate **security + safety issues** that **lead to harm** in order to anticipate and prevent it

Harm modeling

Harm categories - 1

A user sees their post unfairly censored.
This harm is in the category (select one):

- 0% a. Physical injury
- 0% b. Emotional or psychological injury
- 0% c. Opportunity loss
- 0% d. Economic loss
- 0% e. Dignity loss
- ✅ 0% f. Liberty loss
- 0% g. Privacy loss
- 0% h. Environmental impact
- 0% i. Manipulation
- 0% j. Social detriment

URL: ttpoll.eu
Session ID: cs290

Harm categories - 2

A fitness app leaks GPS location data on social media.
This harm is in the category (select one):

- 0% a. Physical injury
- 0% b. Emotional or psychological injury
- 0% c. Opportunity loss
- 0% d. Economic loss
- 0% e. Dignity loss
- 0% f. Liberty loss
- ✅ 0% g. Privacy loss
- 0% h. Environmental impact
- 0% i. Manipulation
- 0% j. Social detriment

URL: ttpoll.eu
Session ID: cs290

Harm categories - 3

Online ads lead a compulsive shopper to additional purchases.
This harm is in the category (select one):

	0%	a. Physical injury
?	0%	b. Emotional or psychological injury
	0%	c. Opportunity loss
?	0%	d. Economic loss
	0%	e. Dignity loss
	0%	f. Liberty loss
	0%	g. Privacy loss
	0%	h. Environmental impact
?	0%	i. Manipulation
	0%	j. Social detriment

Difficult to categorize:

- A human is harmed, we can extrapolate that there is psychological damage
- There is also financial damage for the person, but we are not allocating resources
- There is manipulation of behavior, we can say it harms social systems but it's a bit of a stretch (impact on citizenry unclear)

URL: ttpoll.eu
Session ID: cs290

Harm categories - 4

A recruitment software indirectly discriminates based on people's name. This harm is in the category (select one):

- 0% a. Physical injury
- 0% b. Emotional or psychological injury
- ✅ 0% c. Opportunity loss
- 0% d. Economic loss
- 0% e. Dignity loss
- 0% f. Liberty loss
- 0% g. Privacy loss
- 0% h. Environmental impact
- 0% i. Manipulation
- 0% j. Social detriment

URL: ttpoll.eu
Session ID: cs290

Harm categories - 5

The results of an image search engine for “Nurse” show only women. This harm is in the category (select one):

- 0% a. Physical injury
- 0% b. Emotional or psychological injury
- 0% c. Opportunity loss
- 0% d. Economic loss
- 0% e. Dignity loss
- 0% f. Liberty loss
- 0% g. Privacy loss
- 0% h. Environmental impact
- 0% i. Manipulation
- ✅ 0% j. Social detriment

URL: ttpoll.eu
Session ID: cs290

Instructions

Read the <u>“Smart home technologies” scenario (2nd one)</u>
*Use the “*Affect-Display Textile Garment*” scenario for later practice.*

Individually, fill out the harm table from the template
Use the description of the categories from the cheatsheet

Share with your neighbor:
- Did you identify the same harms?
- Did you classify harms in the same way?

Overall debriefing of the strategy

Determining the type of harm:

1. What is harmed (**category**)
2. What is the **type** of harm in this category

It may not always be evident to classify some types of harms:

- Some harms may **fall into several categories**
- Some categories of harms **overlap**

Argumentation is important!

Not all categories apply to each case!

We should do this **for different scenarios!**
(including not creating the product)

What’s next?

We start Safety 2!

Tomorrow, Tuesday 23: notebook on content *recommendation*

By Monday 29:

- Watch **videos 2.1 to 2.5** + do the **quizzes**
- Finish the notebook
  (and any other leftover from previous weeks)

On Monday 29:

- Interactive questions on the theory
- Work on the **case studies together in class**