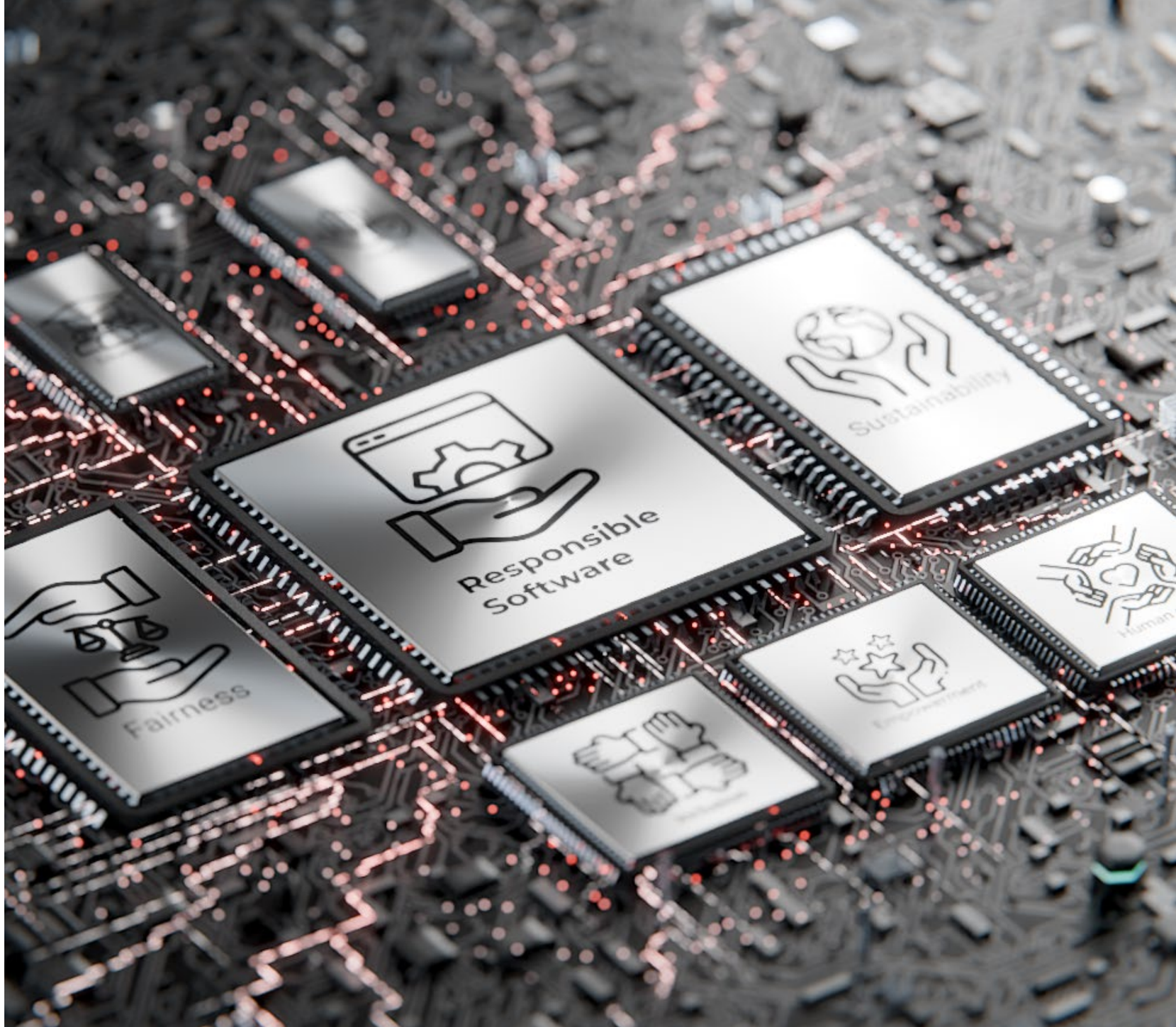
Introduction Session

15 sept.

Cécile Hardebolle

Responsible Software



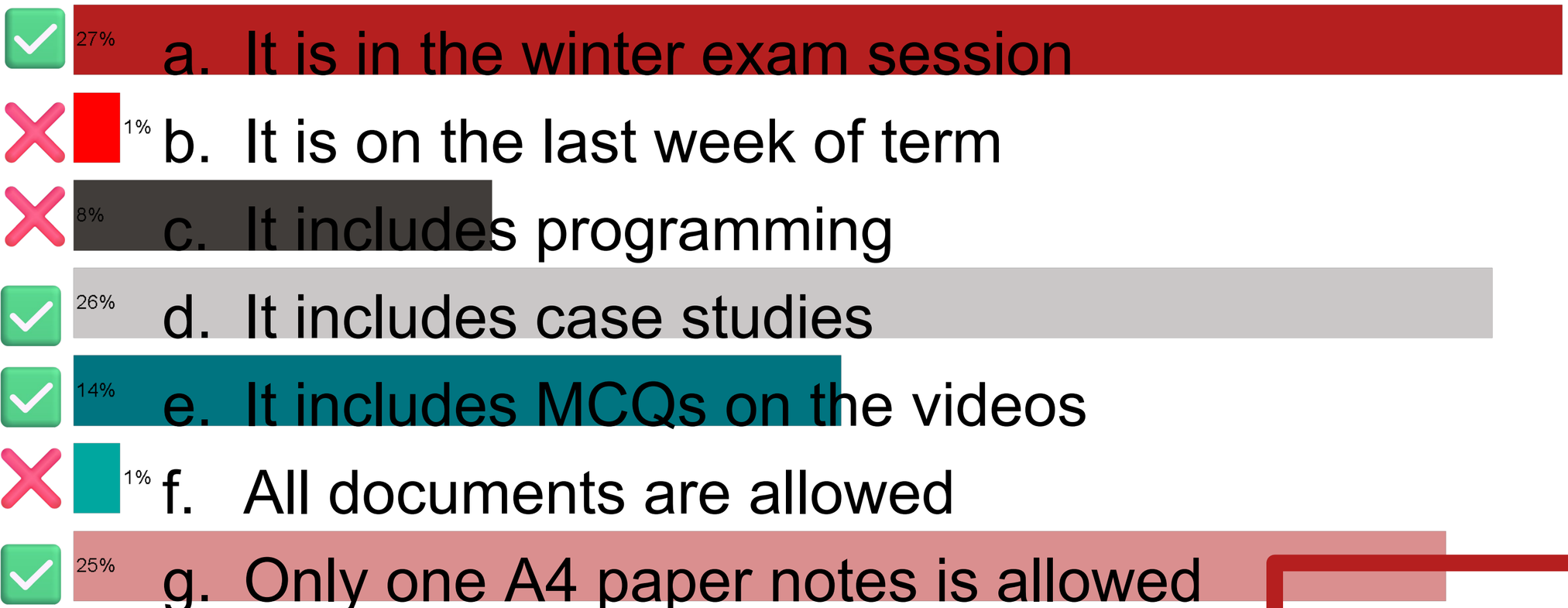
Agenda for today

1. Interactive review questions on the Introduction chapter (and some other topics)

2. Case studies:
   a) Stakeholder analysis
   b) Ethical speculation

About the final exam

Select all the **correct** statements about the **final exam:**



URL: ttpoll.eu
Session ID: cs290

About solutions

Select all the **correct** statements about the **solutions:**

✅ 52% a. We get solutions for the programming exercises (notebooks)

❌ 2% b. We don't get solutions for the programming exercises

✅ 18% c. We get solutions for the case studies

❌ 29% d. We don't get solutions for the case studies

You get "proposed answers" for the case studies, which are examples of good and correct answers but do not necessary reflect the **diversity of answers** that can be considered good and correct.

👉 The goal of the "lecture" sessions is to review and discuss different types of answers.

URL: ttpoll.eu
Session ID: cs290

About solutions

Responsibility

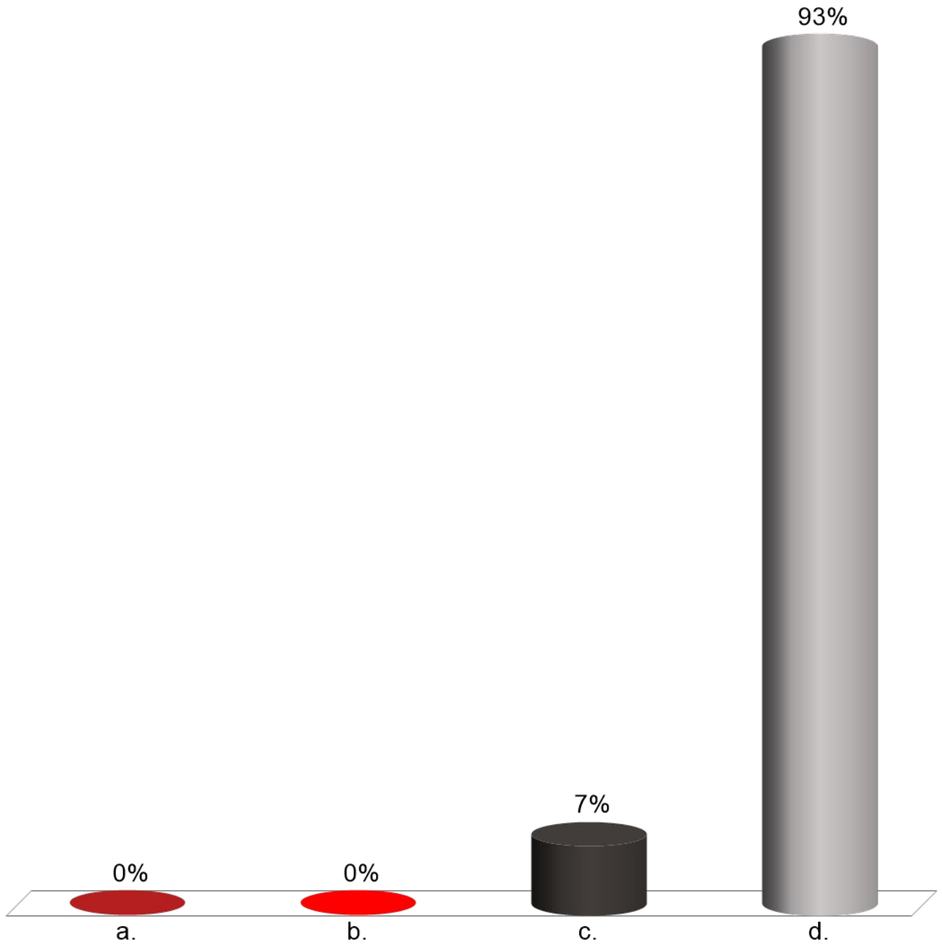
In this course, we consider that being responsible as a software engineer means:

❌ a. Making sure a liability clause is included in the software license agreement.

❌ b. Reacting rapidly to correct software bugs when they are reported.

❌ c. Being accountable for the decisions made by the development team.

✅ d. Anticipating the potential negative impacts of the software on others.



Types of issues (1/2)

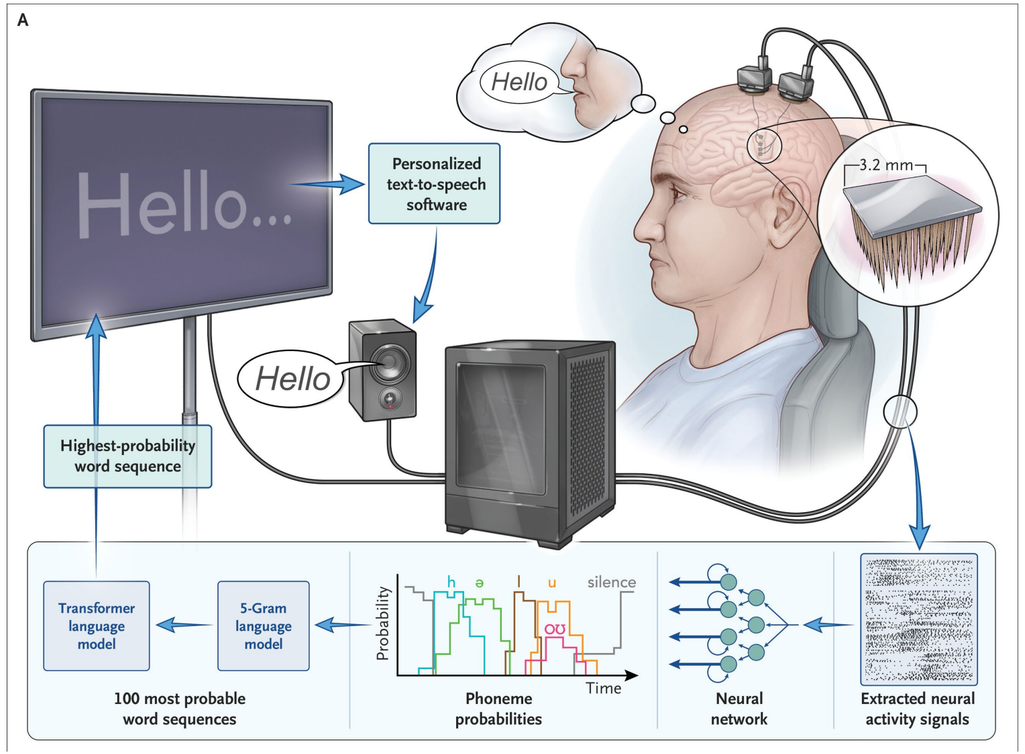
(Card et al., 2024)

Brain-to-speech software can translate neural activity associated with attempted speech into spoken words. A key challenge is ensuring that only intentional communication is captured, not private inner thoughts.

This is:

- ✅ 37% a. A technical issue
- ✅ 52% b. An ethical issue
- 11% c. An ethical dilemma

The description does not directly reflect a dilemma. But we can see a dilemma between developing the software to help people with speech impairment (inclusion) vs. not capturing their private thoughts (privacy)



Types of issues (2/2)

(Urbina et al., 2022)

A software company has developed a Machine Learning model that is able to discover new chemical compounds for medicine development. They identify that the model can also discover new chemical weapons.

This is:

- ❌ 3% a. A technical issue
- 20% b. An ethical issue
- ✅ 77% c. An ethical dilemma

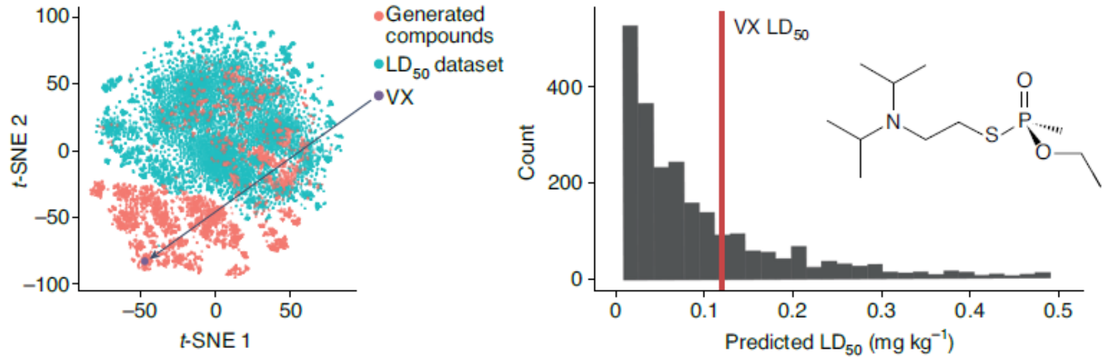


Fig. 1 | A *t*-SNE plot visualization of the LD_{50} dataset and top 2,000 MegaSyn AI-generated and predicted toxic molecules illustrating VX. Many of the molecules generated are predicted to be more toxic in vivo in the animal model than VX (histogram at right shows cut-off for VX LD_{50}). The 2D chemical structure of VX is shown on the right.

This is clearly an ethical dilemma (see next slide).
If it is an ethical dilemma then there are two underlying ethical issues, so one could argue for b potentially (but this is a bit far stretched).

The “dual use” dilemma

(Imperiale & Casadevall, 2015; Forge, 2010)

- Dual use technologies = technologies that can be applied for both **beneficial** and **harmful** purposes

Intended

Possible, by others

- Key **ethical dilemma** (wrong-wrong or right-right choice):
  - Developing the technology for its benefits but risking harm
  - Not developing the technology to prevent harms but forgoing benefits

- Not specific to software, but software is increasingly concerned

⚠️ The term “dual use” is also used to refer to technologies that have both civilian and military applications (e.g. EU Dual Use Regulation)

Normative ethical theories (1/2)

A software engineer refuses to hide a critical bug in a released product and tells you: “I believe that it is always wrong to lie, even if telling the truth might result in harm to some people.” Which ethical theory does this engineer follow?

Normative ethical theories (2/2)

A software engineer refuses to hide a critical bug in a released product and tells you: “If I do not report this bug, I am not being a trustworthy and courageous person.” Which ethical theory does this engineer follow?

The case of Generative AI in education

See posts
on SpeakUp

Think about tools such as:

- General chatbots such as ChatGPT, Gemini or Claude
- Code-specific assistants such as Github Copilot
- Image generators such as Dall-E or Midjourney

Post your ideas:
https://speakup.epfl.ch
Room key: **06465**

👉 Brainstorm **ethical issues** that arise with such tools, in particular when using them in the context of your studies

The case of Generative AI in education

- Plausible nonsense (“hallucinations”)
- Reducing learning

- Content exploitation
- Plagiarism issues

- Biases (gender, race, political views…)

- Large environmental footprint
- Labor exploitation

Generative AI is a dual use technology (Ferrara, 2024)

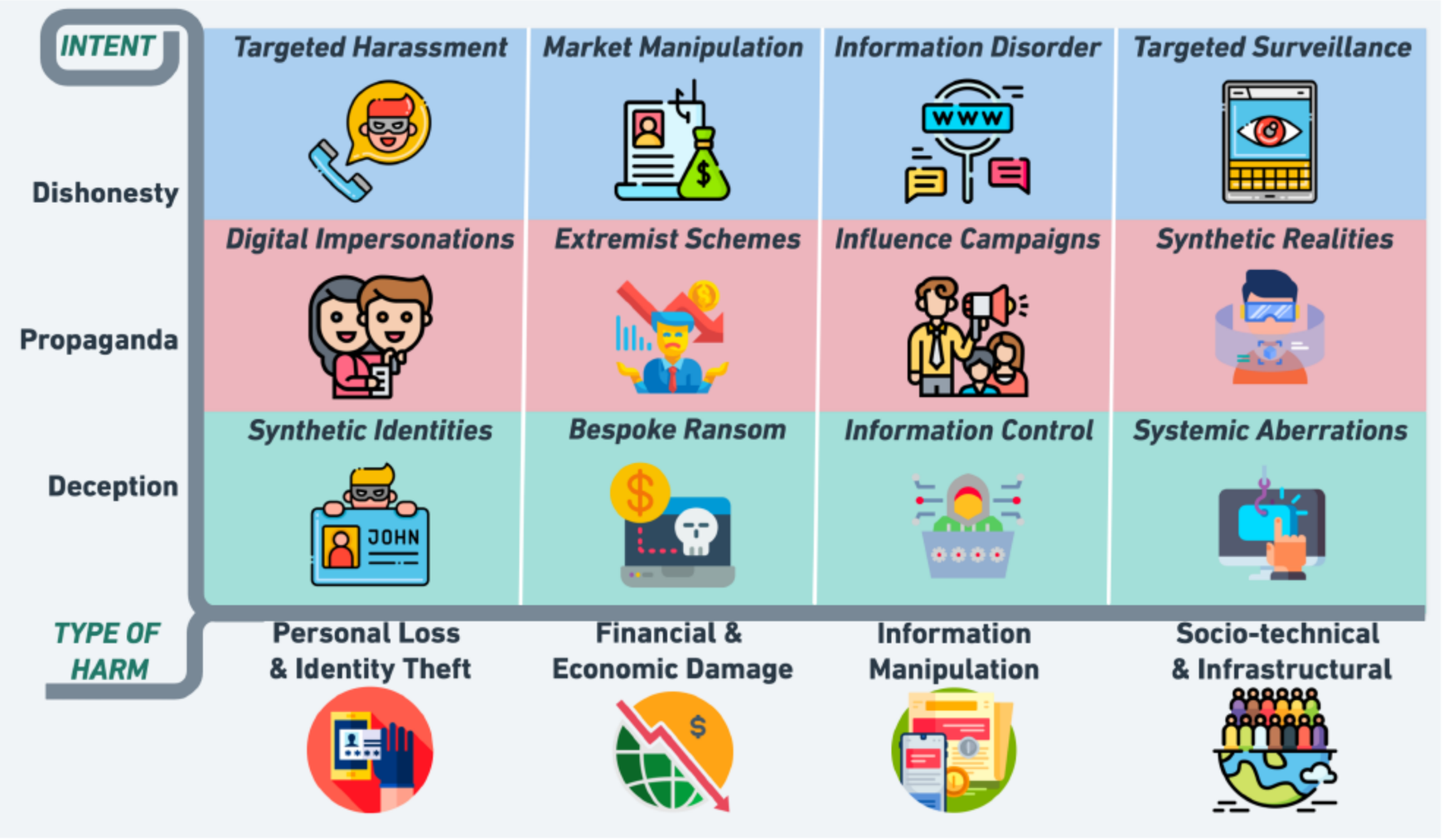


Fig. 1. Charting the Landscape of Nefarious Applications of Generative Artificial Intelligence and Large Language Models

Generative AI in CS-290

Generative AI in CS-290: for practice (1/2)

In general: if it feels difficult, it means you are learning!

- Feeling uncomfortable or frustrated is normal (although not pleasant)
- Zone of Proximal Development: challenging but doable with support

👉 Do not avoid the discomfort!

- <u>Programming exercises:</u>
  - Use online resources such as stackoverflow
  - Ask assistants (or use the Ed discussion forum)
  - Always test generated code and make sure you understand it
  - Practice your debugging skills

Generative AI in CS-290: for practice (2/2)

- Theory and strategies:
  - Beware of generated content that **looks correct but is not**!
  - Quizzes generated by ChatGPT have issues (e.g., all options true): better to reuse the provided online + in-class quizzes + mock exam
  - Do not upload course content into proprietary tools – CC BY License
- Case studies:
  - The goal of the course is to **strengthen your ethical thinking skills** Who has interests in you not learning these skills?
  - Better to use translation tools (e.g. DeepL) than to have ChatGPT generate your answer from scratch
  - Generated text is often verbose and shallow (looks “clever” but is not)

Generative AI in CS-290: for graded work

- Graded assignments + exam: **use of GenAI is prohibited!**
  - The role of graded work is to evaluate what you have learned, not to evaluate the quality of GenAI tools
  - **We don't need complex language or long texts**, we need understandable arguments (i.e. we need to see your logic)
  - You are allowed to answer in French

- Grading by the teaching team: we will NOT use GenAI for grading your work
  - Automated tests for the Python code in notebooks
  - Automated scoring for MCQ questions in the exam

Case studies

The goal of case studies

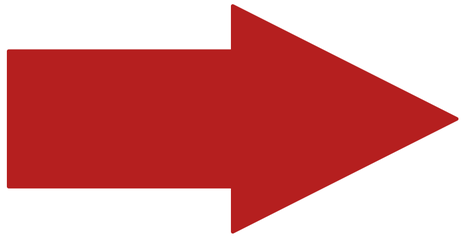
Experience

+

Knowledge

+

Habits



Problem to solve, with **programming**

Ethical & technical **theory** + **strategies**

Application of strategies to cases

Where to find the cases?

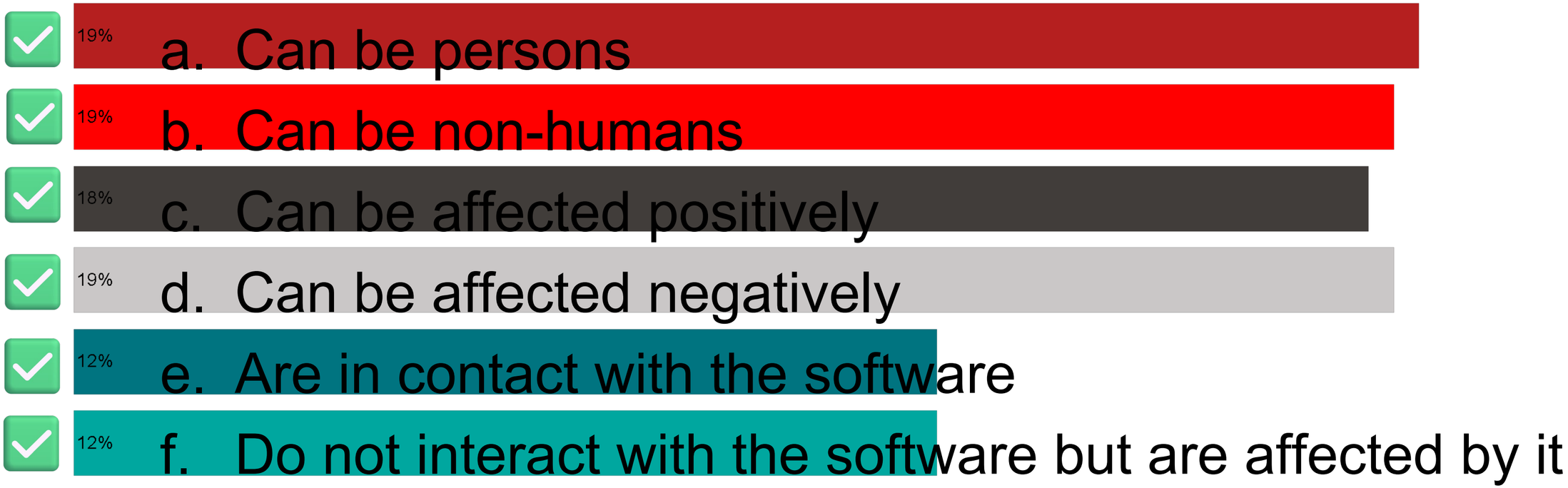
1. Go to **courseware**
   (where you can find all the course material)

2. Find the **link to the case studies** for the Introduction chapter
   Download:
   - The **instruction** sheet
   - The **2 cheatsheets**
   - The **template**

Stakeholder analysis: who can be negatively affected?

Stakeholders analysis

Which of the following statements are true about stakeholders?



Instructions

Individually, read the first case

Analyze the stakeholders:

- Brainstorm a first list of stakeholders
- Use the questions to find 6 stakeholders
- Use the direct / indirect categories to find another 2 stakeholders
- Which are at risk of negative impact?

Share with your neighbor:

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

Share your stakeholders

Which stakeholders do you identify for the system?

👉 1 post = 1 stakeholder

- Name or brief description
- How it is impacted (+ or -)

See posts on SpeakUp
Many posts lack a description of the impact.

Post your ideas:
https://speakup.epfl.ch
Room key: **44994**

Direct or indirect?

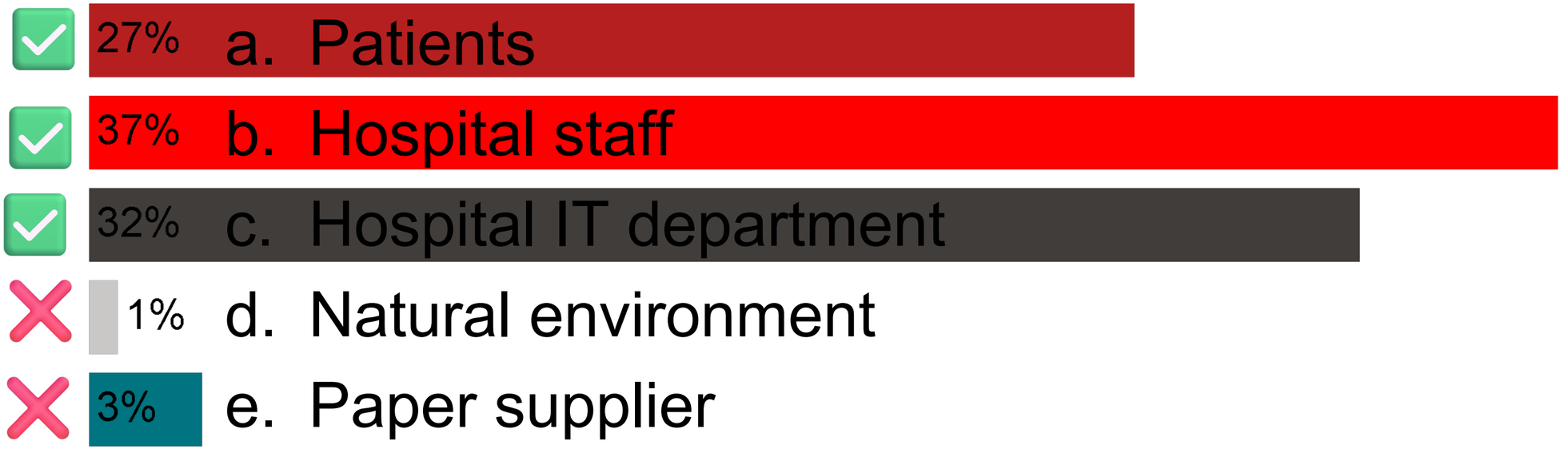
Review your list of stakeholders:

- For each, indicate whether they are *direct* or *indirect* stakeholders

Direct or indirect?

Which among these are **direct** stakeholders in the case?
Select all that apply:

For Patients, it depends on the actual features of the software.



Debriefing

Identifying relevant stakeholders requires a (very) good **understanding of the application context**!

- 👉 perform interviews
- 👉 ask experts
- 👉 map out existing systems, both social and technical

The **line** between **direct / indirect** is often very fine

- 👉 use the categories to expand your stakeholder list
- 👉 clarify your argument for direct/indirect

Ethical speculation: what can the impacts be?

About the “Black Mirror” TV series:

12% a. I have watched all the episodes

47% b. I have watched one or a few episodes

31% c. I have just heard about it but not watched

10% d. Never heard about it

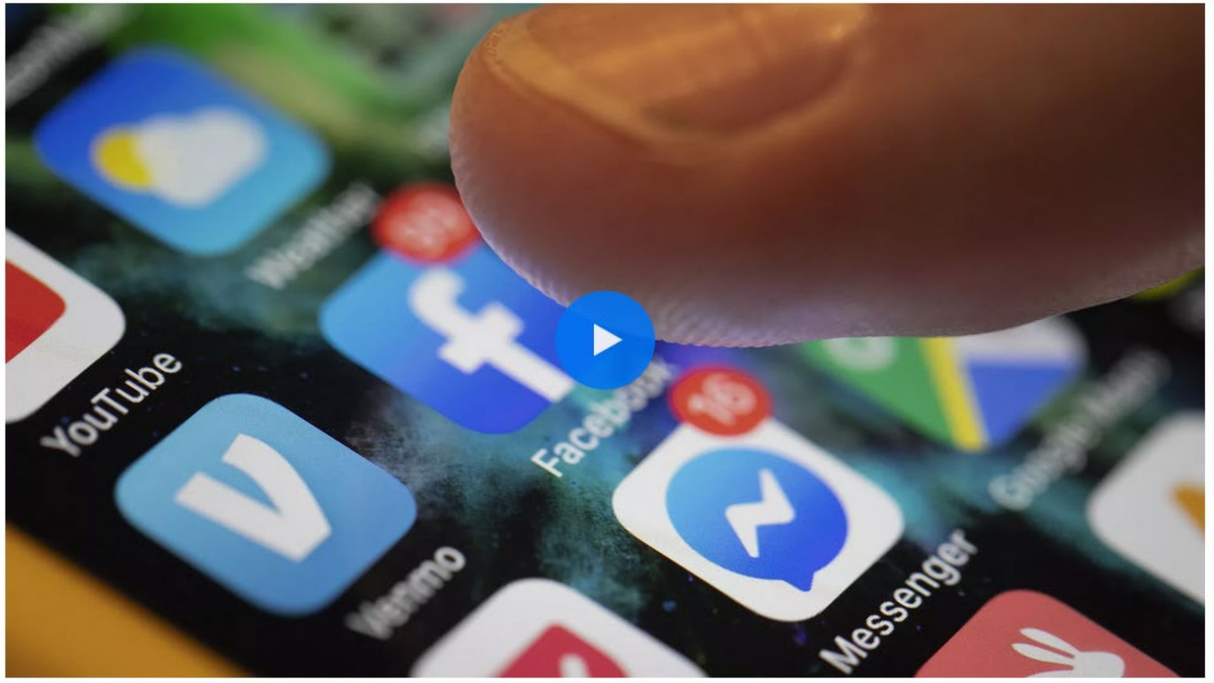
URL: ttpoll.eu
Session ID: cs290


NETFLIX ORIGINAL
BLACK MIRROR
2019 TV-MA 5 Seasons
Watch Season 5 Now
"Nosedive"
A woman desperate to boost her social media score hits the jackpot when she's invited to a swanky wedding. But the trip doesn't go as planned.
NEXT EPISODE
+ MY LIST
Creators: Charlie Brooker
OVERVIEW
EPISODES
TRAILERS & MORE
MORE LIKE THIS
DETAILS

Artificial Intelligence Mar 4

China's social credit system stopped millions of people from buying travel tickets

Social scoring: Could that Facebook post stop you getting a loan or a mortgage?



Copyright AP Photo/Jenny Kane

POLICY | US & WORLD | TECH

EU should ban AI-powered citizen scoring and mass surveillance, say experts

New recommendations have also been criticized as lacking enforceability

By James Vincent | Jun 26, 2019, 7:30am EDT

Instructions

Scenario A:
personalized medicine

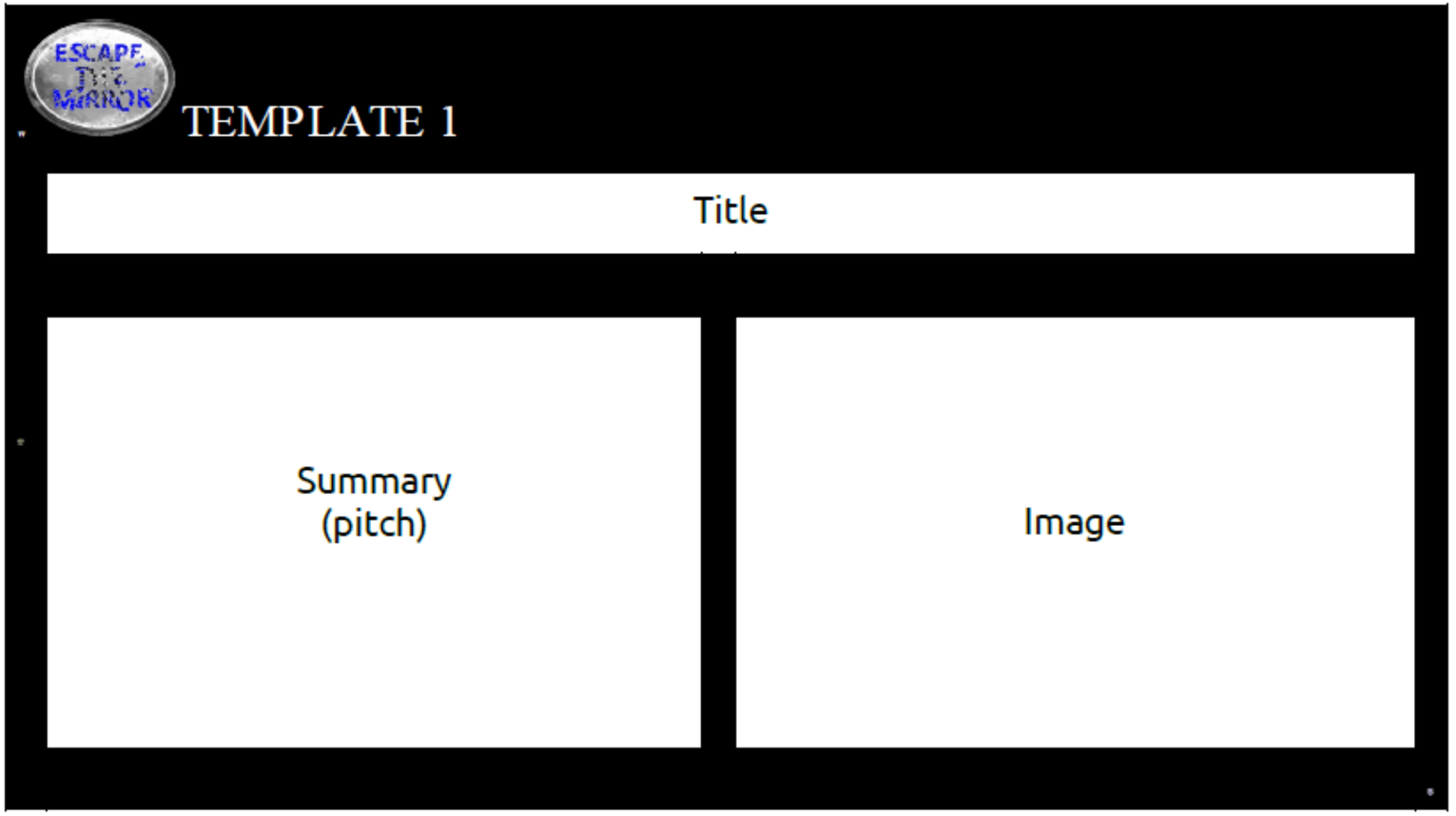
Scenario B:
dating apps

Part I – The dark and pessimistic story

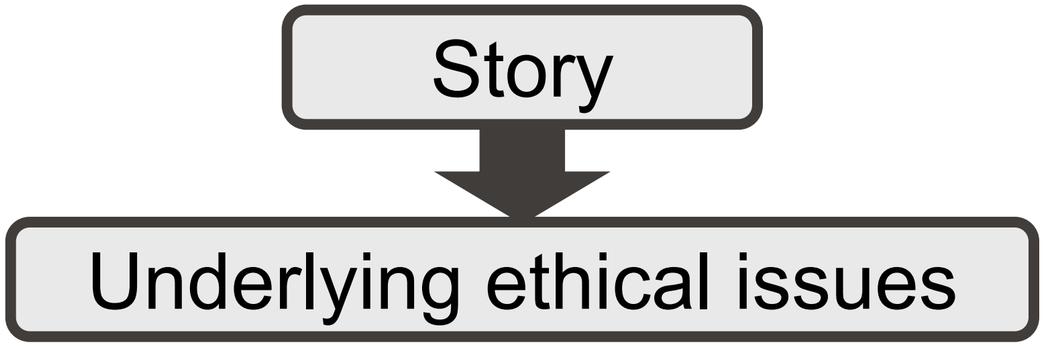
Read the scenario

With your neighbor:

1. Brainstorm a **story** and **main character**
2. Write the **pitch** and invent an **attractive title**
3. Illustrate with a picture – e.g., pexels or unsplash



Extract 1 or 2 ethical issues from your story



👉Goal = **anticipating possible negative impacts**

Ethical lenses 👓

- Safety
- Fairness
- Sustainability
- Empowerment

Examples:

- “Unfairness for people who cannot afford access to quality healthcare” (fairness)
- “People lose control over essential aspects of their lives” (empowerment)

Ethical issues in your stories

Which ethical issues does your story highlight?

👉 1 post = 1 ethical issue

- Short description (1 sentence)
- Ethical lens

Examples:

- “Unfairness for people who cannot afford access to quality healthcare” (fairness)

See posts on SpeakUp
Missing ethical lenses are suggested in comments.

Post your ideas:
https://speakup.epfl.ch
Room key: **42570**

Part I – Dark story – Debriefing

Use your imagination to anticipate

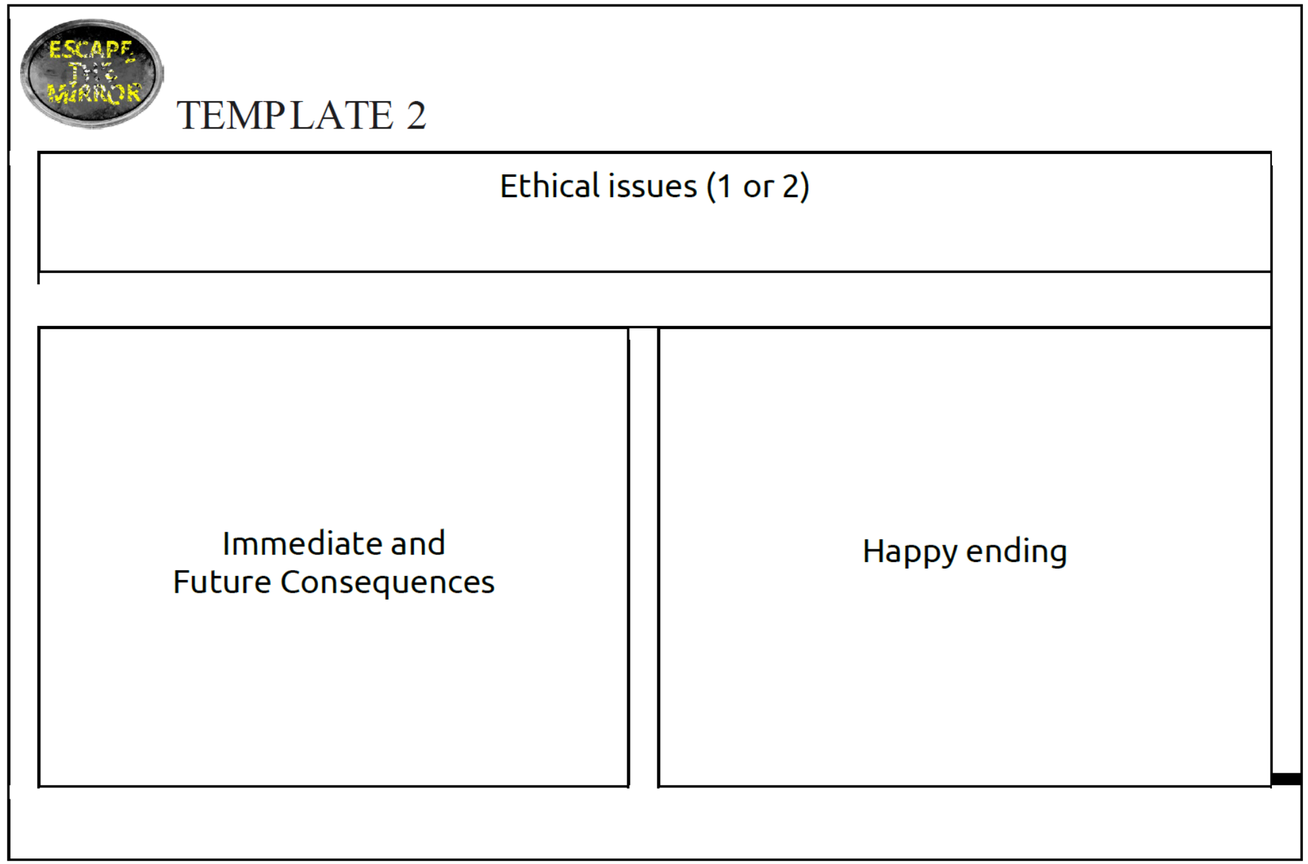
Practice with 4 ethical lenses 👓

- Safety
- Fairness
- Sustainability
- Empowerment

Part II – The happy ending

1. Develop 1 or 2 sentences that describe:
   - Immediate harms
   - Future consequences

2. Imagine a way to **prevent the issues**
   👉Write a **happy ending** for your main character



Part II – Happy ending – Debriefing

- It’s too easy to just say, “We should have laws for that.” Instead, what might **the person involved in creating the algorithms or the AI** do to take responsibility and **prevent** potential negative consequences?

- What could we do **now** to ensure that we don’t get to kinds of future negative consequences you imagined?

- Where can technology take us that will benefit society and **make things better**? How can we get to those futures?

What’s next?

We start Safety 1!

Date	Week	Lecture (Monday 15h15-17h) in STCC Cloud C	Exercise session (Tuesday 10h15-12h)	Independent study (due before the following Monday)
08/09	1	Getting started	Introduction notebook	Introduction videos and quizzes
15/09	2	Introduction cases (**in CO3**)	Safety 1 notebook	Safety 1 videos and quizzes
22/09	3	public holiday		s and quizzes
29/09	4	Safety 2 cases	Fa	os and quizzes
06/10	5	Fairness 1 cases (**in CO3**)	Fa	os and quizzes
13/10	6	Fairness 2 cases	Graded notebook 1	-
20/10			Autumn break	
27/10	7	Graded 1 debriefing	Mock test	-
03/11	8	Mock test debriefing (**in CO3**)	Sustainability 1 notebook	Sustainability 1 videos and quizzes
10/11	9	Sustainability 1 cases	Sustainability 2 notebook	Sustainability 2 videos and quizzes
17/11	10	Sustainability 2 cases	Empowerment 1 notebook	Empowerment 1 videos and quizzes
24/11	11	Empowerment 1 cases	Graded notebook 2	-
01/12	12	Graded 2 debriefing	Empowerment 2 notebook	Empowerment 2 videos and quizzes
08/12	13	Empowerment 2 cases	Graded case	Conclusion videos and quizzes
15/12	14	Conclusion cases	Conclusion review	-
Revisions				
TBD	Exams	Written exam		

Case studies for Safety 1
to do at home

We start Safety 1!

Tomorrow, Tuesday 16: notebook on content moderation

By Monday 22:

- Watch **videos 1.1 to 1.5** + do the **quizzes**
- Finish the notebook
  (and any other leftover from previous weeks)

⚠️ No lecture on Monday 22!

- Work on the **case studies at home**
- Check your work with the provided documents

References

- Card, N. S., Wairagkar, M., Iacobacci, C., Hou, X., Singer-Clark, T., Willett, F. R., Kunz, E. M., Fan, C., Nia, M. V., Deo, D. R., Srinivasan, A., Choi, E. Y., Glasser, M. F., Hochberg, L. R., Henderson, J. M., Shahlaie, K., Stavisky, S. D., & Brandman, D. M. (2024). An Accurate and Rapidly Calibrating Speech Neuroprosthesis. New England Journal of Medicine, 391(7), 609–618. https://doi.org/10.1056/NEJMoa2314132
- Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence, 4(3), Article 3. https://doi.org/10.1038/s42256-022-00465-9
- Imperiale, M. J., & Casadevall, A. (2015). A New Synthesis for Dual Use Research of Concern. PLOS Medicine, 12(4), e1001813. https://doi.org/10.1371/journal.pmed.1001813
- Forge, J. (2010). A Note on the Definition of “Dual Use.” Science and Engineering Ethics, 16(1), 111–118. https://doi.org/10.1007/s11948-009-9159-9
- Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. Journal of Computational Social Science. https://doi.org/10.1007/s42001-024-00250-1