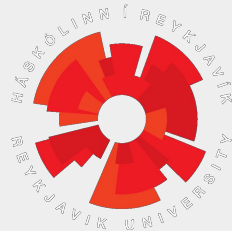


AI, Ethics, and Society

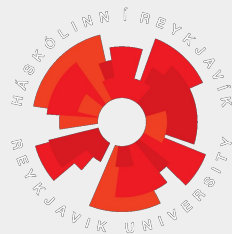
23.09.2026



Recap

Principles of AI Ethics

- Transparency
- Justice and Fairness
- Safety
- Responsibility
- Privacy
- Proportionality



Recap

Things to keep in mind:

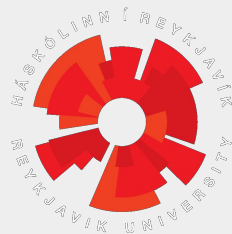
AI (as in mainstream AI) amplifies biases.

No-one is unbiased.

Explanations are crucial to:

- Find these biases
- Better understand safety concerns

Transparency is a major issues. ANNs are opaque and often black-box solutions.



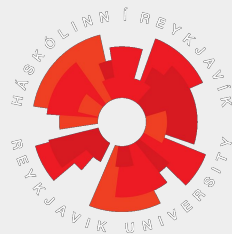
Questions

- How about the user, do they need to be careful when using AI?
- A user that tries hard to get an AI to do something it shouldn't - should they be responsible for what happens?

My answer:

I do not think that the assignment of responsibility is as easy as simply pointing to the developers. Misuse is a major issue. However, the companies should make it as hard as possible for users to misuse the product.

An analogy: A company building fireworks is under very close observation from regulatory bodies. Because they can do harm if misused.

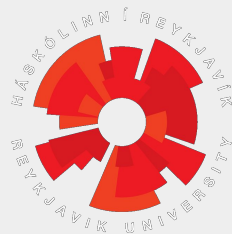


Questions

More training data vs. privacy concerns. How should developers decide?

My answer:

Consent! Consent! Consent! If I need more data from a (e.g.) minority group. Just ask them. Explain to them the effects of their decision (in both directions!). You shouldn't simply take their data without their consent.



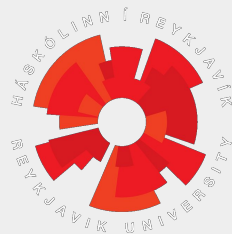
Questions

Regarding Explanations & Transparency:

- How can we know that the explanation actually represents the decision-making process?
- Isn't transparency the most fundamental of these since everything else depends on it?
- How much explanation is necessary so that we can justify using it?

My answer(s):

Explanations must be based on the architecture, design, and representation of the AI. At least until we have justified trust that the AI is actually expressing what it is doing. Once such an explanation exists, we would have made a major step towards a lot of issues with AI. This is why it is so heavily researched. The problem is just that AI develops faster and gets larger at a rate at which research falls behind.

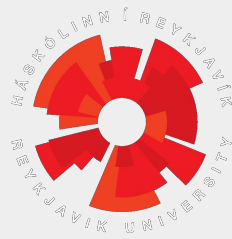


Questions

If an AI system makes decisions that are unpredictable by the developers, how can they be responsible. I.e., is AI still just a tool if it is too complex to predict?

My answer:

Let's look at some other engineering fields and see what we can learn from them.



Who is responsible?

Tacoma Narrows Bridge (1940)

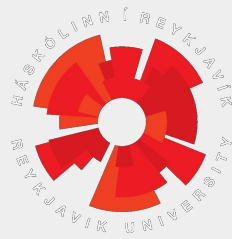
Very slender, elegant bridge designed using the accepted theory of the day.

That theory only considered static wind loads. Nobody had modeled the interaction between wind and a flexible structure, yet. (Now called aerostatic flutter)

Four months after opening, a moderate wind set the deck twisting until it tore itself apart.

The bridge was more slender than anything built before, beyond the range where experience applied.

Is choosing to operate outside tested territory itself a responsible or irresponsible act?



Who is responsible?

Comet airplane (1959)

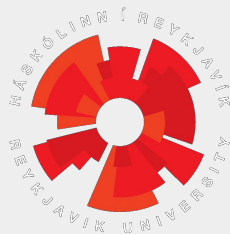
World's first jet airliner. Flew so high the cabin had to be pressurized.

Each flight was one cycle of in- and deflation.

Metal fatigue was poorly understood at that time.

Two comets broke apart.

The company even did fatigue testing in advance. Just the wrong sort. They tested static pressure instead of dynamic fatigue.



Who is responsible?

Malpasset Dam (1959)

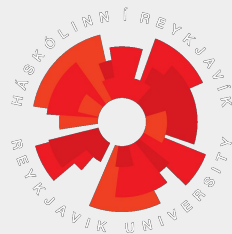
Thin arch dam built by André Coyne.

The concrete actually held. It was a problem with the rock foundation.

Water pressure built up in fissures in the rock under one side, compression made the rock less permeable, thus trapping water instead of letting it drain.

Rock mechanics barely existed as a field at that time.

A wedge of rock slid out, the dam burst and over 400 people died.



Who is responsible?

Vajont Dam (1963)

Looks similar to the previous, but is very different.

A huge landslide fell into the reservoir and sent a wave over the dam that killed around 2000 people.

Geologists had warned about the unstable mountainside. Even earlier creeping movement was registered.

Details were uncertain but danger was known.

Verdicts by the courts

What do you think?

Quick poll.

Left: Should they have been held responsible?

Right: Were they held responsible?

Who is responsible?

Poll ended | 4 questions | 12 of 13 (92%) participated

1. Do you think the designer of the Tacoma Narrows Bridge was held responsible? (Single choice)

12/12 (100%) answered



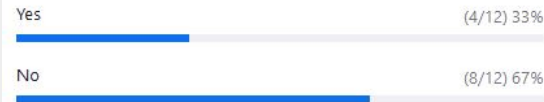
2. Do you think the designer of the Comet ariplane was held responsible? (Single choice)

11/12 (91%) answered



3. Do you think the designer of the Malpasset Dam (rock failure) was held responsible? (Single choice)

12/12 (100%) answered



4. Do you think the designer of the Vajont Dam (landslide) was held responsible? (Single choice)

12/12 (100%) answered



Who is responsible?

Poll ended | 4 questions | 12 of 13 (92%) participated

1. Do you think the designer of the Tacoma Narrows Bridge was held responsible? (Single choice)

12/12 (100%) answered



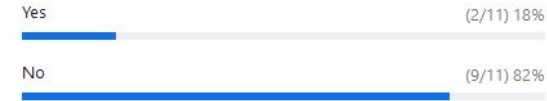
2. Do you think the designer of the Comet ariplane was held responsible? (Single choice)

12/12 (100%) answered



3. Do you think the designer of the Malpasset Dam (rock failure) was held responsible? (Single choice)

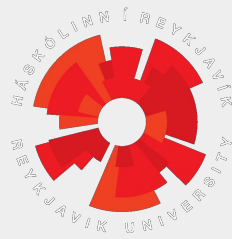
11/12 (91%) answered



4. Do you think the designer of the Vajont Dam (landslide) was held responsible? (Single choice)

11/12 (91%) answered





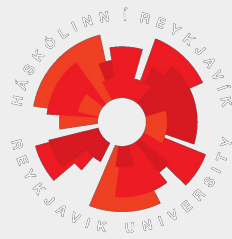
Verdicts by the courts

Tacoma Narrows Bridge (1940)

Nobody was prosecuted or found liable for the design. Investigation was technical, not legal.

Comet airplane (1959)

No one was prosecuted or found negligent. One analysis showed that they even went well beyond the certification requirements in trying to make the aircraft safe. The inquiry's first recommendation was simply: More research! The head of the Air Registration Board even stated that knowing what he knew then, he would not have acted differently.

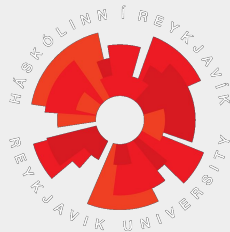


Verdicts by the courts

Malpasset Dam (1959)

No misconduct was found by the Court of Cassation. The Council d'État's final ruling cited geological faults that had been “neither detected nor suspected” during surveying.

Nevertheless, the owner of the dam was held responsible for damages without anyone found at fault.



Verdicts by the courts

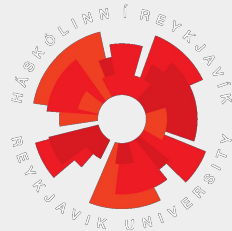
Vajont Dam (1963)

Conviction, because the danger was foreseeable.

Not just because of the measured creeping movement, but because the danger itself was foreseeable.

It was ruled that this was a disaster of negligence, not a natural disaster.

The United nations later classified the disaster as one of the foremost environmental catastrophes caused by human and scientific error.



How can we translate this to AI?

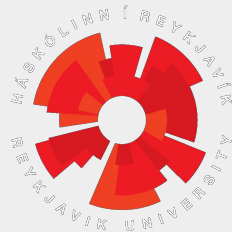
We will look closer at the risks of AI next week.

Uuk et al. (2024) wrote a paper, for example, called **A Taxonomy of Systemic Risks from General-Purpose AI**.

They define 13 categories of systemic risks and 50 contributing sources.

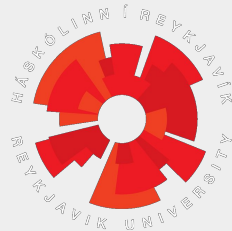
Ranging from environmental harm and structural discrimination all the way to government failure and loss of control.

The risk categories are mapped and published!



Ethics-by-Design & Participatory Ethics

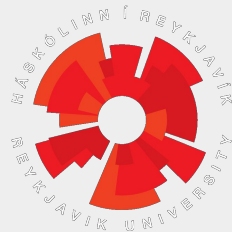
23.09.2026



From Retrospective to Prospective Ethics

Traditional ethics in technology is often reactive: Something goes wrong →
Reflection → Fix

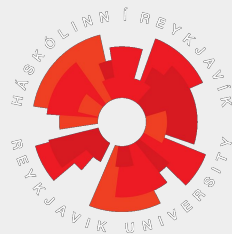
Ethics-by-Design and Participatory Ethics aim to change this: Ethics becomes part
of the initial design and development process.



From Retrospective to Prospective Ethics

Why is this shift necessary?

- AI systems can scale quickly and affect many people before harm is visible.
- Bias and unfairness are often baked in before deployment.
- Retrofitting ethics after deployment is often too late (and too expensive).
- Ethics-by-Design and Participatory Ethics helps to prevent problems rather than solving them afterward.



Ethics-by-Design

What is it?

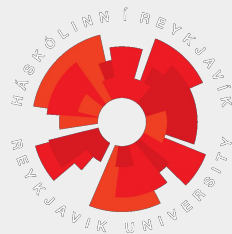
A proactive design approach that seeks to embed ethical values into AI systems from the start. In terms of both architecture and use context.

Purpose:

Ensure that AI reflects human values and legal/social norms, even in ambiguous or high-risk contexts.

Main Features:

Design with ethical principles in mind (fairness, accountability, etc.)



Ethics-by-Design

- Consider intended and unintended consequences.
- Translate ethical values into technical system constraints.
- Use ethics as a requirement, not an add-on.
- Implementation examples:
 - Limit data collection to what is necessary (privacy-by-design).
 - Create interfaces that support informed decision-making.
 - Allow users to challenge or explain outcomes (e.g., explainability modules).

How could such an implementation look like?

Let's take a look at the second point: Interface for decision making.

Ethics-by-Design

Example:

Consent dashboard

Gives the user the ability to revoke and extend consent at all times.

Mitigates privacy issues.

Must be easily accessible and in the best case in simple language.

The screenshot shows a consent dashboard for 'Some AI Company'. At the top, there's a purple header with the company logo and name. A prominent purple banner contains the text 'Help us build a smarter, safer AI for everyone!' and a yellow 'Accept all & continue' button. Below the banner, the dashboard is divided into two main sections. The left section, titled 'Advanced settings (not recommended)', lists several toggle switches for user preferences: 'Don't exclude interaction content from quality programs', 'Limit use of content you provide', 'Opt out of visual experience improvements', 'Restrict audio personalization', 'Don't share feedback with trusted reviewers', 'Disable signal-based enhancements', and 'Exclude developer content from innovation'. The right section, titled 'Legitimate interest', lists purposes for data processing: 'Develop and improve models', 'Measure product performance', and 'Research with selected partners', each with an 'Object' link. At the bottom right, there are two buttons: a purple 'Accept all' button and a white 'Save my choices' button.

Some AI Company

Privacy Choices

Help us build a smarter, safer AI for everyone!

Some AI Company gets better every day thanks to people like you. By sharing, you help improve answers for millions of users worldwide.

★ 98% of members choose to share. Join them!

[Hide options](#)

Advanced settings (not recommended)

Changing these settings may reduce the quality of Some AI Company for you and others.

- Don't exclude interaction content from quality programs** ☒
Includes content, context and related information.
- Limit use of content you provide** ☒
Applies to eligible content only.
- Opt out of visual experience improvements** ☒
Media may still be processed for core functionality.
- Restrict audio personalization** ☒
Voice features may become unavailable.
- Don't share feedback with trusted reviewers** ☒
Helps our quality team.
- Disable signal-based enhancements** ☒
Technical data is always collected for security.
- Exclude developer content from innovation** ☒
May affect code suggestions.

Legitimate interest

We and our 214 partners process data on the basis of legitimate interest. You may object to each purpose individually.

- Develop and improve models** [Objection under review](#)
- Measure product performance** [Object](#)
- Research with selected partners** [Object](#)

Data retention

Data is retained for as long as necessary to provide, maintain and improve our services and those of our affiliates, in accordance with applicable law.

[Accept all](#)

[Save my choices](#)

Created by Claude Opus 5.5

Ethics-by-Design

Example:

Consent dashboard.

Gives the user the ability to revoke and extend consent at all times.

Mitigates privacy issues.

Must be easily accessible and in the best case in simple language.

PRIVACY · DATA & AI TRAINING

Your data, your call.

Choose which kinds of data may be used to train and improve AI models, and for what. Nothing is shared for training unless it is switched on here.

Download consent record

Stop all sharing

Categories shared

4 of 7

Active uses across categories

9

New data categories

Off until you opt in

Data categories

Select a category to see how it's used

Conversations

Messages you send and the replies you receive.

Used for: training new models, safety & misuse research, improving this product



Uploaded files

Documents, spreadsheets and PDFs you attach.

Not used for any AI training



Images SENSITIVE

Photos and screenshots you upload.

Not used for any AI training



Voice SENSITIVE

Audio from voice mode and its transcripts.

Not used for any AI training



Feedback

Ratings and written comments on answers.

Used for: safety & misuse research, improving this product, human review



Usage signals

Which features you use and how sessions flow.

Used for: improving this product



Code

Snippets and repositories you share.

Used for: training new models, improving this product



Conversations

Shared

WHAT IT MAY BE USED FOR

☒ Training new models

Included in datasets used to train future AI models.

☒ Safety & misuse research

Used to detect harmful patterns and make models safer.

☒ Improving this product

Aggregated to fix bugs and improve features.

☐ Human review

May be read by trained, access-controlled staff.

KEEP IN TRAINING DATASETS FOR

30 days

12 months

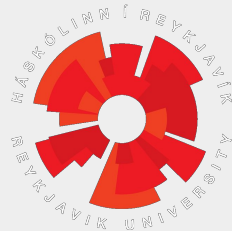
Until I withdraw

What withdrawing means

Switching off stops this data from entering future training runs. Models that have already been trained with it are not retrained, but your data is removed from the datasets we keep.

[Request removal from existing datasets](#)

Created by Claude Opus 5.5



Participatory Ethic

Tries to provide an answer to the question “Who gets to decide what 'ethical' means?”

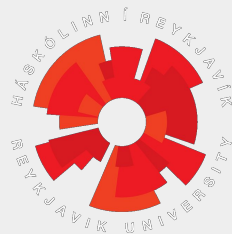
Participatory ethics says: Not just developers or companies, but users, communities, and generally those affected.

What is it?

A process that involves stakeholders and affected parties directly in identifying ethical goals and risks in system design.

Motivation:

Developers can't predict all social impacts



Participatory Ethics

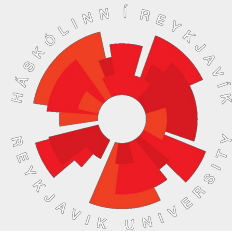
Some users are more vulnerable than others.

Inclusion improves both ethics and usability.

Promote democratic legitimacy of AI deployment.

Participation Methods:

- Co-design sessions
- Public consultation forums
- Focus groups or structured interviews.



Participatory Ethics

Challenges:

Risk of tokenism (including people but not listening)

Conflicting values: fairness for one group may mean exclusion for another

Time and resource-intensive

Not all voices are equally loud or equally heard.