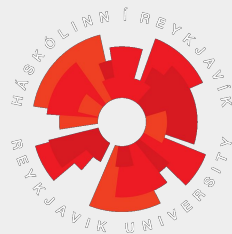


AI, Ethics, and Society

30.09.2026

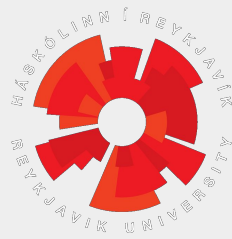


Quick feedback

Did you enjoy the in-class assignment?

Do you think this helped you in better understanding the sort of issues that emerge when different affected parties need to be considered?

Would you recommend me doing such an exercise again next year?



Recap - Where we left off last week

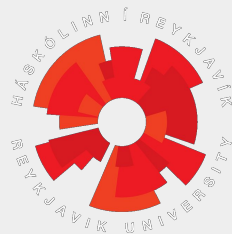
We went through 4 engineering examples from the past where there happened something that was “unpredictable.”

Three of the examples were truly unpredictable in that the failure happened due to unknown unknowns.

In the Isat example, geologists warned about the general risk. The dam was built nevertheless which led to a catastrophe.

There is a difference between “not knowing” and negligence.

Today we look closer into the question of what do we know about the risks of AI.



Recap

Ethics-by-Design:

Building ethical constraints into the design process from the beginning.

See them as system requirements rather than “nice to haves.”

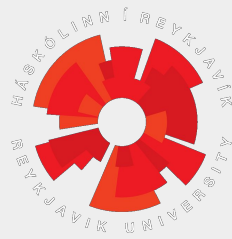
Design around these requirements.

Participatory Ethics:

Bring in people from all affected groups to identify moral tensions.

The group defines what values and morals should be incorporated in the company/process/product.

Can be used to identify the requirements for an Ethics-by-Design approach.



Questions

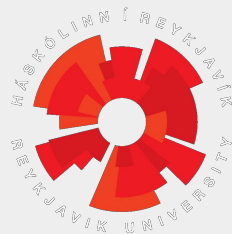
You had a lot of questions about how the courts would rule today, or when is it negligence and when truly unforeseeable.

TBH this is for the courts to decide. I am not a student of law, I really can't tell you how these decisions are made by the courts.

What I can say is, that ignoring the danger signs and risks is usually a bad sign.

Especially when they don't come from individuals, but a whole research group.

Sometimes you hear people say things along the lines of "They are always over cautious" or "You can never make it right for them." (Them being an ethics board, workplace security, etc.). I don't think that's a good argument (and neither do the courts afaik).



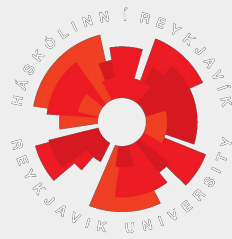
Questions

How can developers know when they have properly applied Ethics-by-Design?

How can developers measure the effectiveness of their Ethics-by-Design approach?

By continuous evaluation. Ethics-by-Design shouldn't be a one of. Continuous evaluation (ideally by an external ethics board or similar) gives you insights about the effectiveness of your approach.

Ethics-by-Design is, in essence, similar to any Risk Mitigation Study you already might know. You need to think of every part of your system, what risks might emerge, and how to mitigate them. And then, continuously verify that you are within your requirement profile.

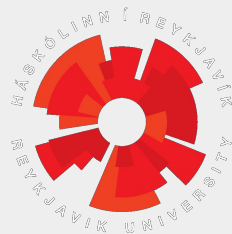


Questions

If a developer needs diverse data to correct a dangerous systemic bias but cannot feasibly acquire individual consent from thousands of people, should they prioritize bias mitigation or strict privacy?

Difficult question. Usually, the “Do no harm” credo is more important than “Doing good”. But what if both sides are harms? There is a third option: Turn of the system. But that has a lot of harms for your employees.

A fourth option would be to go to the press and be honest about these problems. Transparency is another principle and it shouldn't only hold for the AI system, but for the whole company behind it. But yes, this is a hard decision with no easy answers.

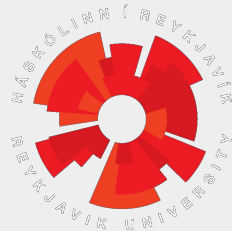


Questions

Are these approaches only available for large corporations due to their costs, etc.?

I don't think so. First of all, the importance of your decisions is directly related to the number of affected parties. So a small company might be more flexible in their choices. Also, Participatory Ethics becomes easier with smaller groups of affected parties.

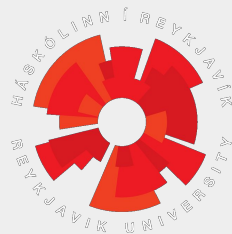
Secondly, there exist external ethics boards you can contract for ethical evaluations.



Questions

How to manage Participation such that affected communities are heard and loud groups not overrepresented?

There are mitigation processes for these sorts of things. But honestly, I think this is a bit outside of the scope of this lecture. If you are really interested, I could do a deep-dive into this, though. I am just not sure if we want to spend our time on this. Let me know what you think.



Questions

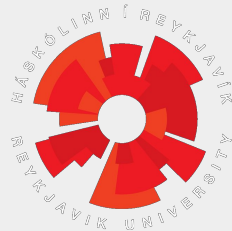
Do you think awareness for Ethics-by-Design and Participatory Ethics is increasing?

Yes, I do think so. In public bodies, it is already well established.

In any funding application I have ever written or read, there was an Ethical Considerations section.

People are becoming more aware of unethical practices (at least from my experience).

Nevertheless: Still a long way to go. Most important is the education of people regarding these issues (that's why we are here, I guess).

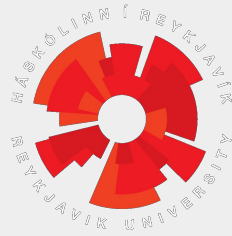


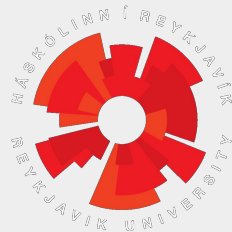
Questions

Some questions about the implementation in regulations, the usage of risk categories, etc.

We will look at that some time later in the course when we look at the EU AI Act as an example of a regulation that includes ethical considerations.

Let's look at another example for a minute





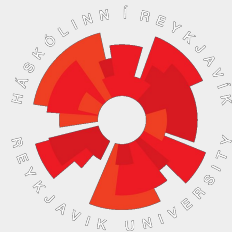
In-Class (or after class discussion)

Hypothetical example:

A university uses AI to flag students at risk of failing a course, based on attendance and grades.

- **Ethics-by-Design**
 - What values would need to be built into this system from the start? What technical or design decisions would need to reflect those values?
- **Participatory Ethics**
 - Who should be consulted or involved during the design phase? Why? Whose perspectives might otherwise be ignored?

What's one ethical concern that would be hard to solve through design alone?

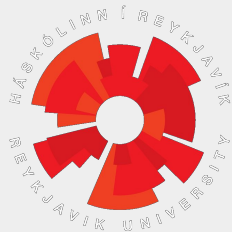


AI in Practice: What Can Go Wrong?

As AI systems move from lab environments to real-world applications, new kinds of ethical and practical problems emerge.

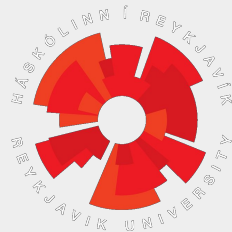
Why does this matter?

- AI systems may work well in testing but fail unpredictably in the real world.
- Ethical risks are often hidden in data, deployment context, or incentives.
- Failures can scale quickly and impact real lives
 - Discrimination, safety issues, loss of trust.
- Understanding how and why AI goes wrong is essential to preventing future harm.



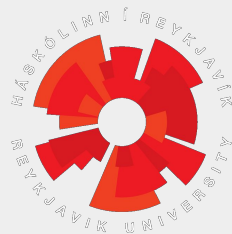
What can go wrong in Practice?

- **Bias and Discrimination**
 - Biased training data leads to biased outputs. Replicates and reinforces social inequalities
 - Example: Hiring tools, predictive policing
- **Lack of Transparency**
 - Users don't understand or cannot challenge AI decisions. Systems act as “black boxes”
 - Example: Credit scoring, medical diagnosis
- **Overreliance / Automation Bias**
 - People trust AI even when it is clearly wrong or misaligned with context
 - Example: GPS directions, autopilot
- **Function Creep**
 - AI used for one purpose expands silently into others (e.g., law enforcement use of commercial data)
 - Example: Smart speakers, surveillance



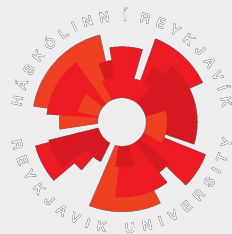
What can go wrong in Practice?

- **Data Privacy Violations**
 - Personal data collected without proper consent. Data being reused or sold
 - Example: Smart homes, mental health apps
- **Unsafe Deployment**
 - AI deployed in real-world contexts without sufficient testing or safeguards
 - Example: Self-driving cars, robotic surgery
- **Feedback Loops**
 - System learns from its own outputs, reinforcing narrow behavior (e.g., filter bubbles)
 - Example: Recommender systems, ad targeting



Why do these Failures happen?

- **Biased Training Data**
 - Data reflects past human biases (e.g., racist policing records, gendered job roles)
- **Lack of Contextual Testing**
 - Systems tested in narrow environments don't generalize (e.g., from private roads to city streets)
- **Misaligned Objectives**
 - Optimizing for engagement, clicks, or efficiency can ignore fairness or well-being
- **No Human Oversight**
 - Systems make decisions without accountability mechanisms or intervention
- **Incentive Misalignment**
 - Companies optimize for speed or profit, not ethics or safety
- **Lack of Regulation or Standards**
 - Little to no legal limits on harmful deployment or poor design



Categories of AI Risk

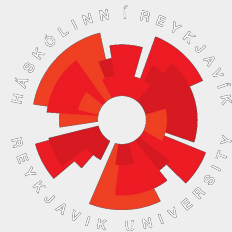
Epistemic risk

- The system is wrong or misleading
- Self-driving car misclassifies child as inanimate object

See “Hallucinations are mathematically inevitable”

“Large language models are increasingly relied upon as sources of information, but their propensity for generating false or misleading statements with high confidence poses risks for users and society.”

- Ghafouri et al. (2025) - Epistemic Integrity in Large Language Models
(<https://arxiv.org/abs/2411.06528>)



Categories of AI Risk

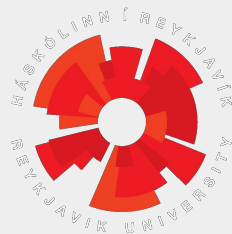
Moral risk

- The system does harm or violates values
- Hiring AI favors privileged groups

1 *Discrimination & toxicity*

1.1 Unfair discrimination and misrepresentation	Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and representation of those groups.
1.2 Exposure to toxic content	AI that exposes users to harmful, abusive, unsafe, or inappropriate content. May involve providing advice or encouraging action. Examples of toxic content include hate speech, violence, extremism, illegal acts, or child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory political speech, or pornography.
1.3 Unequal performance across groups	Accuracy and effectiveness of AI decisions and actions are dependent on group membership, where decisions in AI system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users.

Slattery et al. (2024) - The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence
(<https://arxiv.org/abs/2408.12622>)



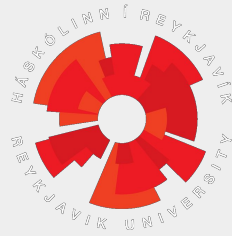
Categories of AI Risk

Political Risk (Power, Manipulation, Structural Bias)

- The system shifts power, often unequally
- Surveillance AI targets protesters. Data sold to authoritarian regimes

“Our findings reveal a complex landscape of potential threats, ranging from environmental harm and structural discrimination to governance failures and loss of control”

Uuk et al. (2024) - A Taxonomy of Systemic Risks from General-Purpose AI
(<https://arxiv.org/abs/2412.07780>)



Categories of AI Risk

Social Risk (Behavioral / Relationship Change, Norms, Polarization)

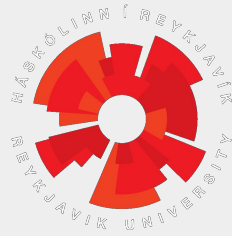
- The system changes relationships or behaviors
- Recommender AI increases polarization and echo chambers

“Domain 5: Human-computer interaction

5.1 Overreliance and unsafe use. Users may come to trust or rely on AI systems beyond their actual capabilities or to anthropomorphize AI system [...]

5.2 Loss of human agency and autonomy. As AI systems become increasingly capable and intelligent, humans may be tempted to delegate many of their decisions and actions to AI (Paes et al., 2023)”

- Slattery et al. (2024) - The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence (<https://arxiv.org/abs/2408.12622>)



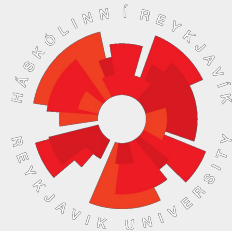
Categories of AI Risk

Legal Risk (Conflict With Law, Lack of Clarity, Compliance)

- Legal Risk (Conflict With Law, Lack of Clarity, Compliance)
- Chatbot offers health advice that violates medical regulations

“6.5 Governance failure. Governance failure refers to the risks and harms that arise when institutional, regulatory, and policy mechanisms fall short of effectively managing and overseeing the development and deployment of AI systems. Several issues make robust AI governance challenging to implement (Nah et al., 2023).”

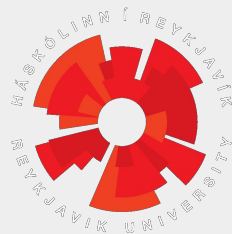
- Slattery et al. (2024) - The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence
(<https://arxiv.org/abs/2408.12622>)



Systemic Risk Categories

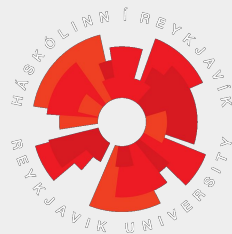
All of the following is taken from Uuk et al. (2024) - A Taxonomy of Systemic Risks from General-Purpose AI

Link: <https://arxiv.org/abs/2412.07780>



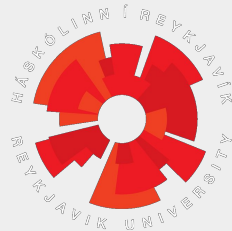
Systemic Risk Categories

- **Control**
 - The risk of AI models and systems acting against human interests due to misalignment, loss of control, or rogue AI scenarios.
- **Democracy**
 - The erosion of democratic processes and public trust in social/political institutions
- **Discrimination**
 - The creation, perpetuation or exacerbation of inequalities and biases at a large-scale.
- **Economy**
 - Economic disruptions ranging from large impacts on the labor market to broader economic changes that could lead to exacerbated wealth inequality, instability in the financial system, labor exploitation or other economic dimensions.



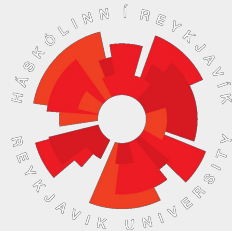
Systemic Risk Categories

- **Environment**
 - The impact of AI on the environment, including risks related to climate change and pollution.
- **Fundamental Rights**
 - The large-scale erosion or violation of fundamental human rights and freedoms.
- **Governance**
 - The complex and rapidly evolving nature of AI makes them inherently difficult to govern effectively, leading to systemic regulatory and oversight failures.
- **Harms to non-humans**
 - Large-scale harms to animals and the development of AI capable of suffering.



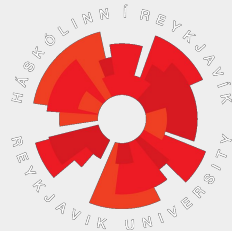
Systemic Risk Categories

- Information
 - Large-scale influence on communication and information systems, and epistemic processes more generally.
- Irreversible change
 - Profound negative long-term changes to social structures, cultural norms, and human relationships that may be difficult or impossible to reverse.
- Power
 - The concentration of military, economic, or political power of entities in possession or control of AI or AI-enabled technologies



Systemic Risk Categories

- **Security**
 - The international and national security threats, including cyber warfare, arms races, and geopolitical instability.
- **Warfare**
 - The dangers of AI amplifying the effectiveness/failures of nuclear, chemical, biological, and radiological weapons.



Takeaways

- AI systems fail not only due to bugs, but also due to ethical blind spots.
- What works in lab conditions can be harmful in society.
- AI failures often reinforce existing social inequalities.
- Most failures are predictable and therefore preventable.
- Ethical foresight is part of responsible design, not idealism.