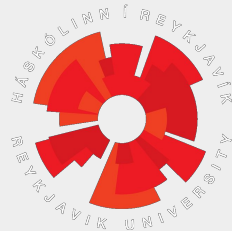


# AI, Ethics, and Society

21.09.2026



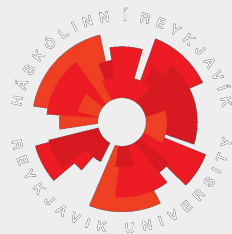
# Recap

## **Agency, Moral Agency, Moral Patiency**

Agency: The functional capability to act toward a goal

Moral agency: Requires additionally intentionality, self-awareness and genuine accountability. A bar that current AI has not cleared.

Moral patiency: being owed moral consideration. Grounded in sentience or interest.

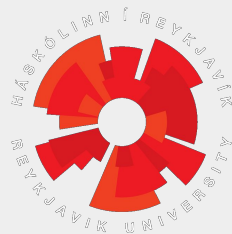


# Recap

## **There is no sharp line (for AI or for us)**

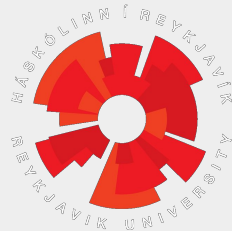
Human age-of-responsibility thresholds (usually age 14 or similar) are pragmatic lines drawn around the development progress of children. Not some sort of proof that morality switches on at a birthday.

No official framework treats real systems as approaching moral personhood.



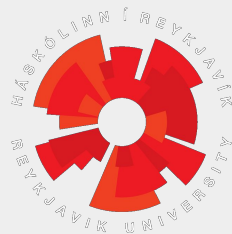
# Questions

1. Where's the line between "can act on its own" and "can be held morally responsible"?
2. If technology is "just a tool," does that still hold once it starts doing things its own developers didn't predict?
3. When ethical frameworks (or the Six Lenses) genuinely conflict, how do you choose?
4. Autonomy: The engineering sense versus the Kantian sense
5. Does the instrumentalist theory of technology still hold for AI?



# Principles of AI Ethics

16.09.2026



# Main Principles of AI Ethics

Transparency

Justice and Fairness

Safety

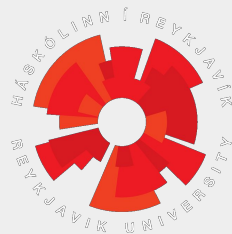
Responsibility

Privacy

Proportionality

Taken from Jobin et al. “The global landscape of AI ethics guidelines.” Nature Machine Intelligence 1.9 (2019)

UNESCO added principle (taken from: UNESCO. “Recommendations on the ethics of Artificial Intelligence” UNESDOC digital library (2022))



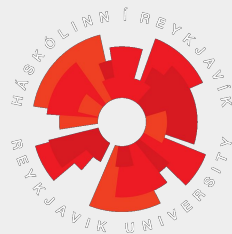
# Transparency

## What is it?

Understandability of decisions/ operations/ plans of systems

## Importance

Building trust, scrutinizing the issues, and minimizing harm



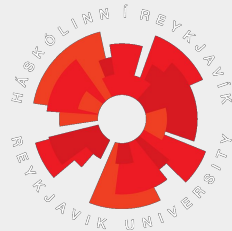
# Transparency

## Issues

- Accountability issues of black-box systems, e.g., LLMs for medicine
- Safety issues: No transparency leads to no improvement, leading to less safety
- Others?

## Implementation

- Knowledge representation and reasoning using explicit causal networks, decision trees, etc.
- Explainability techniques (Post-hoc and real-time)
- Documentation including, for example, (ANNs) open training (and test) data and model types, (Symbolic AI) reasoning types



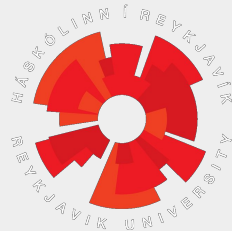
# Transparency

If a public sector or institution makes a decision, the people whose rights have been affected should know whether an AI system plays a role in decision-making.

In such cases, the people have the right to request explanations of why and how the decision has been made.

The institution should provide information and revise the decision if proper explanations do not exist.

The designers whose AI systems affect the rights of other humans should commit to choosing/designing explainable algorithms and systems.



# Justice and Fairness

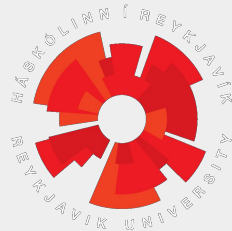
## What is it?

The promotion of equality by avoiding bias and discrimination. This includes:

- Equality in use/ access (e.g., open-source)
- Equality in training/opportunity: No one must be excluded from AI training
- Equality in AI-based judgement: Decision-making based on specific factors, gender, and race is not allowed.

## Importance

Technology must be accessible and useful to everyone.



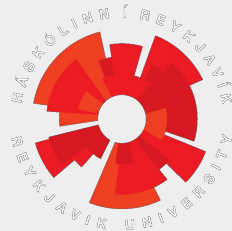
# Justice and Fairness

## Issues

- Conflict of interest between technology providers and users in the competitive AI market.
- Biases in the dataset and data-sensitive algorithms that use these datasets, leading to unequal impacts.

## Implementation

- Following guidelines for data collection
- Reducing the reliance on data-dependent methods
- Bias detection and correction algorithms.



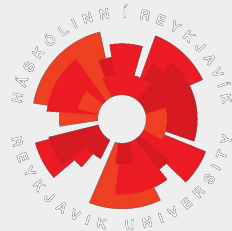
# Safety

## What is it?

Making sure systems do not cause harm to individuals (or society).

## Importance

“Do no harm” is more important than “Doing good”



# Safety

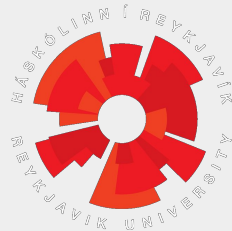
## Issues

Privacy violations, physical harm, or psychological damage, for example:

- Misuse of private information by companies (e.g., facial recognition)
- Physical harm due to malfunctioning or improper use (e.g., accidents in autonomous cars)
- Stress and social anxiety due to interaction with AI systems/ AI companion chatbots.

## Implementation

- Using testing and monitoring techniques
- Anomaly detections
- Safety guidelines for the creation and use of AI systems for specific domains



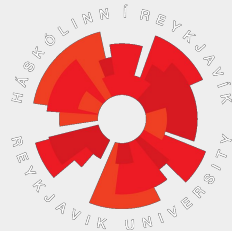
# Responsibility

## **What is it?**

Accountability of actions and decisions

## **Importance**

The need for ethical conduct by AI developers and users



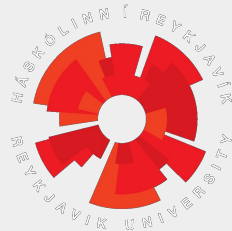
# Responsibility

## Issues

- It is usually not clear who is responsible for the mistakes of AI systems. Ai, designer or user?

## Implementation

- Ethical training regarding the use and development of AI
- Promoting a culture of integrity within AI development teams
- Proper documentation for the system



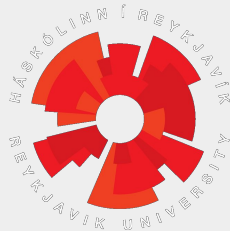
# Privacy

## What is it?

Protection of personal data

## Importance

Protecting the rights of the individual



# Privacy

## Issues

- It is challenging to find the balance between privacy and the need for large datasets for data-driven AI systems development

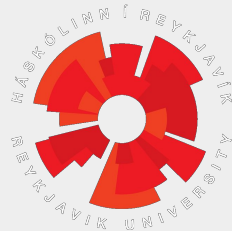
## Implementation

Technical methods like

- Data minimization techniques,
- Privacy-by-design approaches

As well as

- data protection regulations
- Increased public awareness about the privacy rights



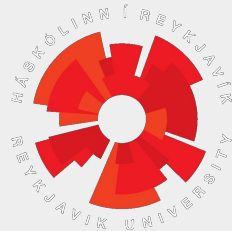
# Proportionality

## What is it?

- AI systems must not be used beyond what is necessary to achieve a legitimate aim.
- Developers and deployers should assess and prevent harms, requiring that any use of AI be appropriately scaled and carefully considered relative to its purpose

## Importance

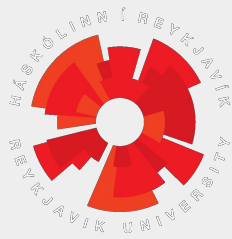
Building trust, scrutinizing the issues, and minimizing harm



# Proportionality

## Importance

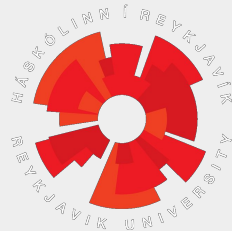
- Helps ensure AI does more good than harm by insisting that risk is weighed and that harm prevention is central.
- Protects individuals, society and rights (e.g., fairness, privacy).
- Helps maintain public trust, avoids overreach (e.g., surveillance, misuse), and ensures ethical limits to what AI can do.



# Proportionality

## Issues

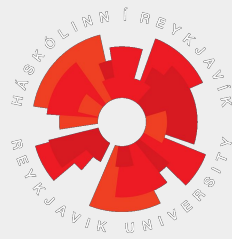
- What counts as “necessary” or “legitimate aim” is often contested.
- Harms can be indirect, delayed, or unseen, making assessment difficult.
- Risk of mission creep: What starts as a minor system may expand beyond original scope
- Sometimes proportionality conflicts with other principles like fairness, or transparency.
- Asymmetric power: Actors with more power may define what is “necessary” in ways that favor themselves.



# Proportionality

## Implementation

- Use risk assessment and harm impact assessment before, during, and after deployment.
- Limit scope: Ensure capabilities are aligned with what is needed.
- Define and document legitimate aims explicitly.
- Ensure there is the ability to stop or scale back systems.
- Build in legal or regulatory constraints to enforce limits.



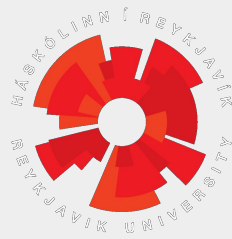
# AI - What are we talking about?!

## Historical Perspective:

In 1950, Turing proposed the idea of building child machines, which are machines that can mimic human children's learning and reasoning behavior.

Minsky (1952) built the first Artificial Neural Net (ANN), called SNARC.

In 1959, Rosenblatt created the Perceptron, which signaled the coming of sub-symbolic systems over half a century before contemporary ANNs. Biological neural networks have inspired the creation of these networks.



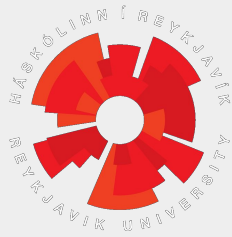
# AI - What are we talking about?!

## Historical Perspective:

In 1950, Turing proposed the idea of building child machines, which are machines that can mimic human children's learning and reasoning behavior.

Minsky (1952) built the first Artificial Neural Net (ANN), called SNARC.

In 1959, Rosenblatt created the Perceptron, which signaled the coming of sub-symbolic systems over half a century before contemporary ANNs. Biological neural networks have inspired the creation of these networks.



# AI - What are we talking about?!

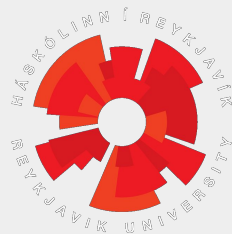
## Historical Perspective:

In 1956, McCarthy and Minsky proposed ideas for building reasoning systems based on logic rules and symbolic representations.

In the 1960s and 1970s, heavy emphasis was put on symbolic AI and rule-based systems such as chess.

The 1970s and 1980s marked the AI winter, where the progress in AI research slowed down.

In the 1990s and 2000s, AI matured in the form of machine learning (ML) algorithms, which are data-driven methods, as opposed to rule-based symbolic systems.



# AI - What are we talking about?!

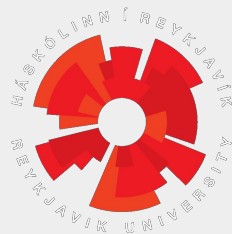
## Contemporary AI: Practical Artificial Neural Networks:

In the 2010s, Artificial Neural Networks (ANNs) required enormous amounts of data and computing power. ANN led to extensive automation in the domains of

- Image recognition: inputs are image data.
- Natural Language Processing: text and audio data

In recent years, ANNs have been used in building large language models (LLMs), e.g., GPT models.

- Training data: texts on the internet.
- How do ANNs get training?
  - Training: feed input data (examples) with known outputs, which tunes the weights (edges between nodes) until it predicts well.
  - Backpropagation: adjust the weights when a wrong prediction is made.



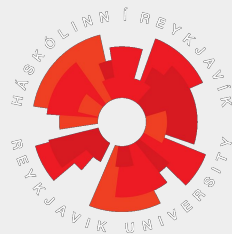
# AI - What are we talking about?!

## **Modern Perspective:**

We now have agentic models, performing massive tasks with little to no oversight.

Gigantic models trained on massive amounts of data.

Chain-of-thought lead to strong advances in the complexity (and length) of tasks that agentic AI can perform.



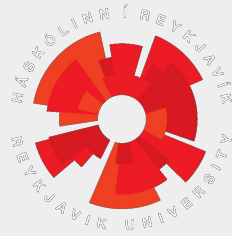
# Ethical Issues of ANNs

## Transparency

ANNs designed for real-world application domains have a large number of inputs and outputs, hidden nodes, and layers, making them black-box predictors/classifiers. (Some examples of domains?)

## Justice and Fairness

ANNs are overly sensitive to training data. If the training data is biased, they amplify the biases through their architecture, making decisions/ classifications/ predictions that reflect horrifying discrimination (e.g., racial and gender biases in what domains?).



# Ethical Issues of ANNs

## **Safety and Security**

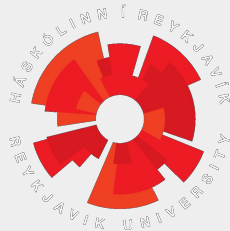
ANNs can not deal with adversarial attacks, where the input test data is designed in such a way that fools the ANN-based AI systems.

ANNs can be used for the creation of deepfakes, which can be used for unethical purposes.

## **Responsibility**

A lack of transparency makes it difficult to understand how the system makes decisions.

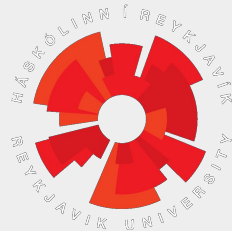
This allows some AI system designers to avoid the consequences of their systems' malfunctions.



# Ethical Issues of ANNs

## Privacy

ANNs require large amounts of training data, allowing ANN developers to collect and access a significant amount of people's private data.



# Can ANNs become more ethical?

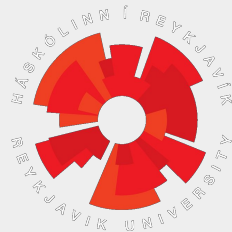
## **Explainable ANNs**

Using explainability techniques to help users and developers understand the decision. But: Only to some extent.

## **Bias mitigation**

Collecting more training data, or bias detection algorithms. However:

- Lack of training data in many domains
- Privacy issues with collecting more and more data.
- Who writes the algorithms? How to ensure that biases are mitigated?



# Can ANNs become more ethical?

## **Safety improvement**

By adding humans in the loop components to retrain models. However:

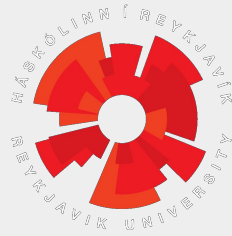
- High human labor cost
- Catastrophic forgetting.

## **Responsibility**

Still hard to make out who is responsible for the decisions made.

## **Privacy**

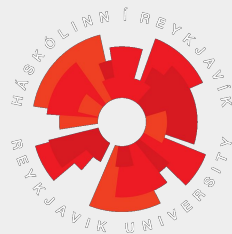
Anonymizing training data (challenging for large datasets).



# Examples

## ChatGPT (OpenAI)

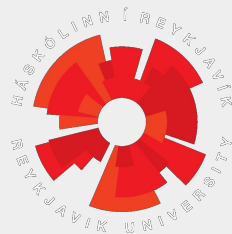
- Current flagship: GPT-5.6 (Sol/Terra variants); predecessor GPT-5.5 launched April 2026.
- Parameter count: unconfirmed. Third-party estimates range wildly — from ~300B–3T on one tracker, to a ~10T claim (later walked back to ~1-2T after methodology issues were found), to Reddit/leak rumors as high as 52T. Treat any specific number as speculation.
- Training data: not disclosed by OpenAI at all for GPT-4/5.x; no credible TB or token-count figure exists publicly.
- Context window: 1,050,000 tokens input, 128,000 tokens output (confirmed, official spec).



# Examples

## Gemini (Google)

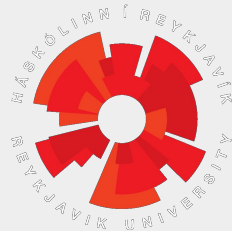
- Current flagship: Gemini 3.1 Pro, built on Gemini 3 Pro.
- Parameter count: undisclosed.  
Estimates range from ~1T total (with only ~50-80B active per token, since it's Mixture-of-Experts) up to a rough "vibe-math" estimate of ~7.5T (confidence interval 1.7T–12.6T). One analyst argues it's the same 1T as Gemini 2.5, with gains coming from better training rather than more parameters.
- Training data: one estimate suggests ~100T tokens seen in pretraining. Unconfirmed.
- Context window: 1,000,000 tokens input, 64,000 tokens output (confirmed, official spec).



# Examples

## Claude (Anthropic)

- Current lineup: Opus 5, Sonnet 5, Haiku 4.5, plus the restricted Fable 5.1.
- Parameter count: undisclosed by Anthropic. One estimate (later revised) put Opus 4.7 at ~4T, subsequently corrected to a more plausible ~1-2T range alongside GPT-5 variants. Still unconfirmed.
- Training data: not disclosed.
- Context window: Opus 5/ Sonnet 5: 1,000,000 tokens input, 128,000 tokens output (confirmed, official spec).



# Ethical Issues

**Transparency:** Very limited and it is unclear how some responses are generated.

**Justice and fairness:** Use of ChatGPT may violate copyrights, etc.

**Safety:** Generating harmful content, misinformation or being misused for unethical purposes.

**Responsibility:** Who is responsible?

**Privacy:** Massive amounts of data. Some training data can be extracted.

See for example [Extracting Training Data from ChatGPT](#)