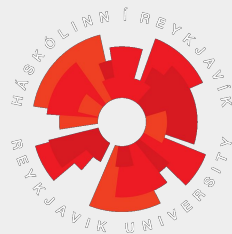


AI, Ethics, and Society

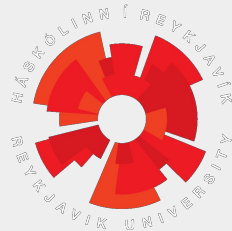
16.09.2026



Recap

Ethics != Morality

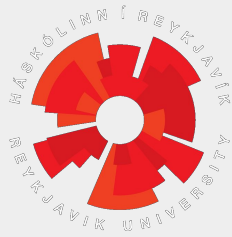
- Morality is the norms a person or a society actually holds
- Ethics is the reasoned, critical analysis of these norms. So “Moral Theories” are, in essence, ethical theories about morality.
- Ethics draws its criteria from an outside theory and reflective testing against strong case judgments.



Recap

The Trolley Problem

- It is a diagnostic, not a policy.
- It isolates variables (numbers, relationships, action vs. inaction) to reveal which principle is driving a judgment.
- It might come up some time to introduce a distinction, but we really want to look at messier cases like AI hiring, content moderation, autonomous vehicles, and so on.

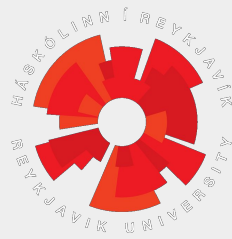


Questions

1. Moral Objectivism: If moral facts exist, how would we ever know?

My answer: We don't. It is an open problem for objectivism.

Another problem is: Even if we knew they existed, how could we encode them into AI training? In theory, we could, but in practice we lack this discovery procedure to identify them. So training will have to approximate shared human judgments rather than an objective truth.



Questions

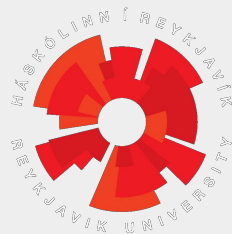
2. Can we trust autonomous AI with ethical decision making?

My answer: Current techniques can train AI to behave such that it follows specified rules (at least mostly). BUT: This is behavioral mimicry, not moral understanding or agency (see today's lecture).

The core unsolved problem is “alignment”. Ensuring what a system actually optimizes for matches what we intend. Including edge cases we didn't foresee.

Trustworthy autonomy requires explainability and robust corrigibility (it reliably accepts correction). Neither is solved yet.

-> Most ethicists would say AI shouldn't be given unsupervised authority over high-stakes moral tradeoffs.



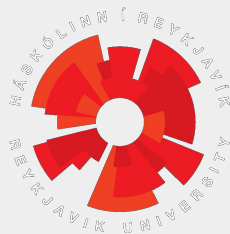
Questions

3. Whose values?

My answer: No single authority currently sets ethical norms for AI. It's a patchwork. Companies, internal principles, regulators (see EU AI Act), government bodies, and public pressure. REAL INSTITUTIONAL GAP!

Regarding multicultural society: Most commonly, the universal baseline (basic rights, safety, non-discrimination) is chosen with culturally configurable norms layered on top. A full separation for different cultures risks giving harmful practices a pass.

This affects not only large scale systems but also at a personal scale (an individual's AI assistant, for example).



Questions

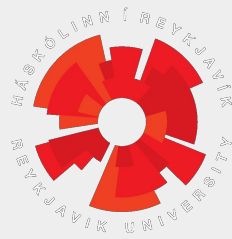
4. Choosing between theories. What to do if they conflict?

My answer: There is no real algorithm for picking the best option.

In practice, different theories are judged for their internal consistency, their fit with strong intuitions, and how they handle edge cases. Hybrid approaches exist that try to keep a theory's core ideas while blocking its worst implications.

Most everyday moral judgment is a plurality of theories. Weighing intentions and outcomes rather than picking a single lens

We will look into this today!



Questions

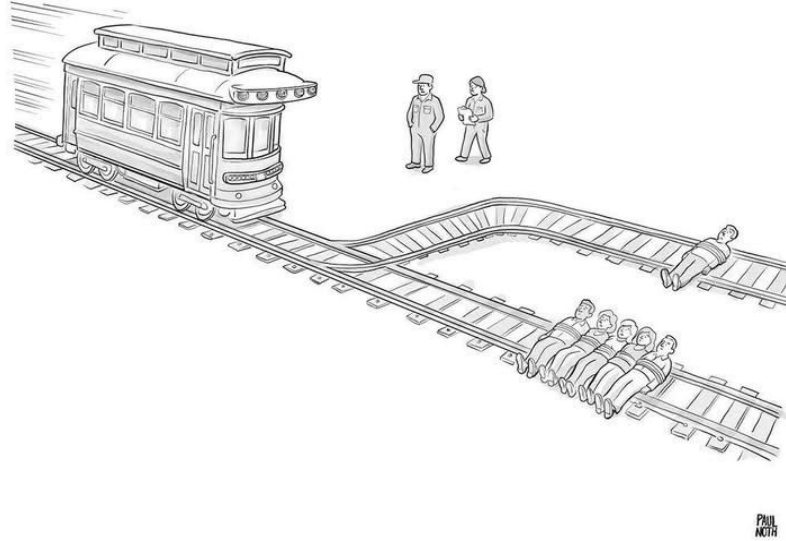
One important summary:

The theories introduced are part of a whole spectrum of different ways to approach ethics. Kant isn't the only deontologist, just probably the most influential.

In reality, we often find a plurality of ideas, intuitions, practices, norms, etc.

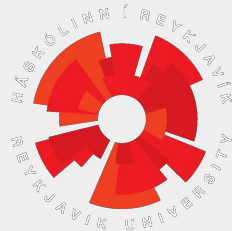
In my experience, especially students of the STEM sciences hope to find a singular answer to the question “What is right or wrong?”. But quite often it doesn't exist.

I am just trying to give you ideas and ways to think about these problems or decisions.



*"Luckily, the A.I. trolley will
decide who to kill all on its own."*

Thanks @Valdís for this! Taken from the New Yorker

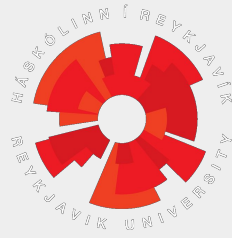


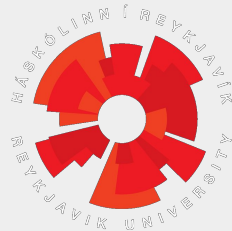
Moral Theories II

16.09.2026

Concepts

Agency?





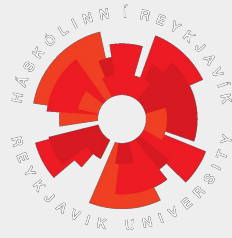
Concepts

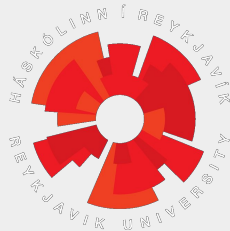
Agency:

- The ability to do something that counts as an action (Himma 2009).
 - Basically doing something.
 - Actions as doings.
- So is simply existing enough? - Not quite.
- Requires a certain mental state

Concepts

Patency?





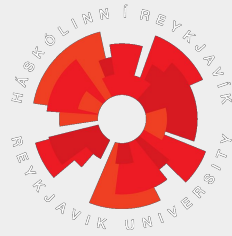
Concepts

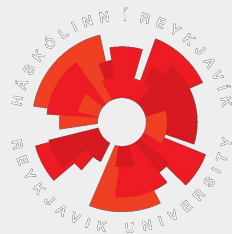
Patience:

- Inanimate objects
- Something that is imposed upon the will of an agent
 - A rock
 - A steering wheel
 - A trolley

Concepts

Moral Agency?





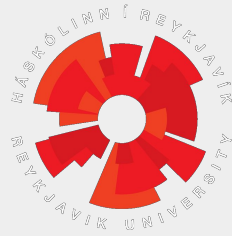
Concepts

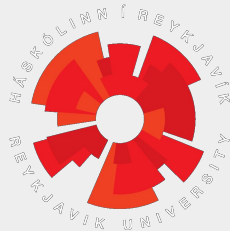
Moral Agency:

- Beings whose behavior is subject to moral requirements
 - Moral obligations
 - Accountability for one's actions
- Agency is a prerequisite for moral agency
- BUT: Not all agents are moral agents

Concepts

Moral Patiency?

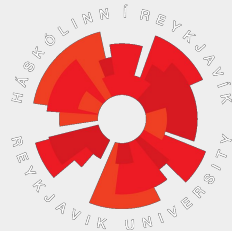




Concepts

Moral Patiency:

- Agents who do not meet the requirement of moral agency
- Someone who is owed at least one duty or obligation (Himma 2009)
 - Newborn infants
 - Animals



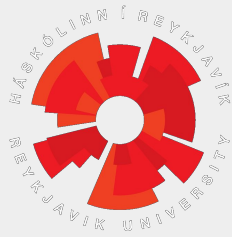
The Instrumentalist Theory of Technology

Technology is a tool for humans. An extension of man.

E.g., the computer, the hammer, or the washing machine.

Any moral violations are clearly the responsibility of the developers or users.

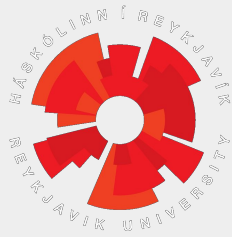
This way, the computer cannot be used as the “scapegoat”.



The Instrumentalist Theory of Technology

Problems:

- Anthropocentric/ Excluding
- Historically, what counts as an agent or moral agent has been prone to change.
- Does it hold that all technology is an extension of man?
 - Is it simply a tool that we use?
 - Does all technology have patience (as assumed under the IT)?
- What about simple machines?
- What about AI?
- Can we extend agency to technology?



Autonomy as a Requirement for Moral Agency

Cognitive/ Operational Autonomy

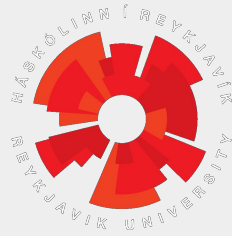
What it is:

The ability of an agent to act and think independently.

The ability to do tasks without interference or help from others or from outside itself.

Implies that the machine “does it alone”.

Refers to the mental (control-) independence of agents - the more independent they are (of their designers, of outside aid, etc.) the more autonomous they are. Systems without it could hardly be considered to have general intelligence.



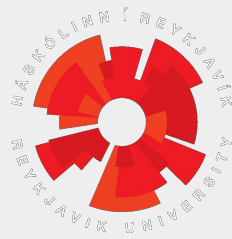
Cognitive/ Operational Autonomy

Structural Autonomy

Refers to the process through which cognitive autonomy is achieved: Motivations, goals and behaviors as dynamically and continuously (re)constructed by the machine as a result of changes in its internal structure.

Constitutive Autonomy

The ability of an agent to maintain its own structure (substrate, control, knowledge) in light of perturbations.



Cognitive/ Operational Autonomy

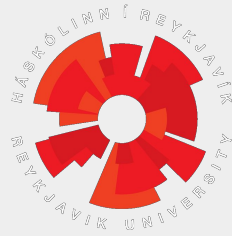
“Complete” Autonomy?

Life and intelligence rely on other systems to some extent. The concept is usually applied in a relative way, for a particular limited set of dimension when systems are compared, as well as same system at two different times or in two different states.

Reliability

Reliability is a desired feature of any useful autonomous system.

An autonomous machine with low reliability has severely compromised utility.
Unreliability that can be predicted is better than unreliability that is unpredictable.



Cognitive/ Operational Autonomy

Predictability

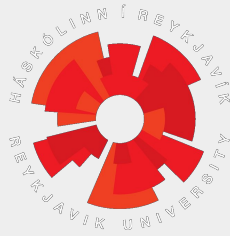
Predictability is another desired feature of any useful autonomous system.

An autonomous machine that is not predictable has severely compromised utility.

Explainability

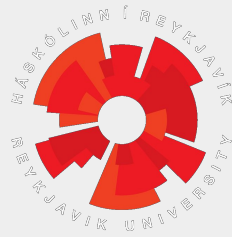
Explainability is a third desired feature of any useful autonomous system.

An autonomous machine whose actions cannot be explained cannot be reliably predicted. Without a reliable prediction, a machine cannot be trusted.



Cognitive/ Operational Autonomy

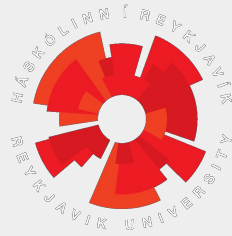
- “In Kant's metaphysics, autonomy refers to the fundamental condition of free will - the capacity of the will to follow moral laws which it gives to itself” (Winner 1977).
- Kant contrasts this view with the concept of heteronomy. The rule of the will by external laws, or the deterministic laws of nature.
- How does “Autonomous Technology” fit into this?
 - When we say technology is autonomous, is it then nonheteronomous, i.e., not governed by external law?



AI as autonomous moral agents?

AI pushes the boundaries of the definition of autonomous moral agents. However, assigning any form of responsibility or obligation to the entity itself is a perplexing and inconceivable endeavor.

- Philosophers and legal scholars still struggle with this question.
- Practical application of AI ethics seems to be taking a different stance.
- Rather than assigning moral responsibility to the AI system, responsibility resides with developers and users of the technology.
- However, ethical challenges are recognized and addressed in the development and implementation process with ethical frameworks.



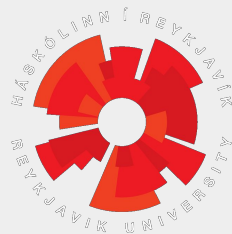
The Six Lenses of Ethical Decision Making

Created as a tool to look at the same decision from six different points of view.

Can be used to find tensions within the decision and its outcomes.

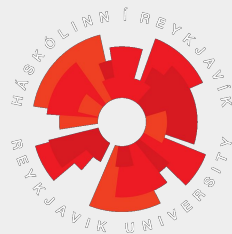
Created by the Santa Clara Markulla Center for Applied Ethics.

Detailed analysis can be found [On the Santa Clara Markulla Center for Applied Ethics website](#)



The Six Lenses of Ethical Decision Making

- The Rights Approach
- The Justice (or Fairness) Approach
- The Utilitarian Approach
- The Common Good Approach
- The Virtue Ethics Approach
- The Care Ethics Approach



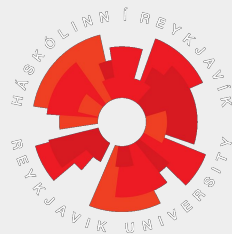
The Rights Approach

Core Idea:

Every individual has moral rights that should be respected, including rights to truth, privacy, freedom, and fairness

Key Question:

Does this action respect the moral rights of everyone involved?



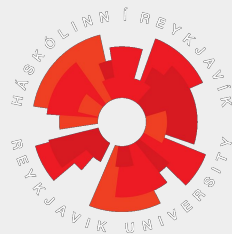
The Rights Approach

Decision Making Focus:

- Avoid violating anyone's inherent rights
- Don't treat people as mere means to an end
- Rooted in deontological ethics (esp. Kant)

Useful when:

- There's a risk of exploitation, coercion, or deception
- You want to ensure informed consent and respect for autonomy



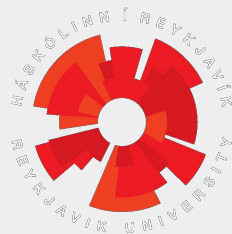
The Justice (or Fairness) Approach

Core Idea:

Ethical actions treat people equally or, if unequally, based on relevant differences (e.g., need, effort, responsibility).

Key Question:

Is this action fair? Are benefits and burdens distributed justly?



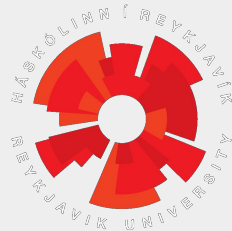
The Justice (or Fairness) Approach

Decision Making Focus:

- Treat similar cases similarly
- Use principles of fairness to evaluate outcomes
- Consider institutional or structural inequalities

Useful when:

- Decisions affect groups differently
- There's potential bias, favoritism, or systemic injustice



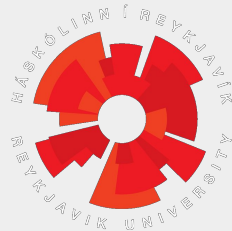
The Utilitarian Approach

Core Idea:

Choose the action that produces the greatest good for the greatest number;
minimize harm, maximize benefit.

Key Question:

What outcome will create the most overall good (or the least harm)?



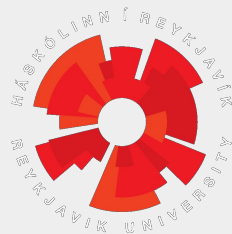
The Utilitarian Approach

Decision Making Focus:

- Weigh consequences for all affected parties
- Often used in public policy or cost-benefit analysis

Useful when:

- You need to evaluate trade-offs or side effects
- Ethical concerns involve resource allocation, risk, or public safety



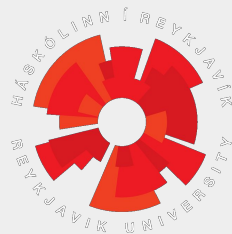
The Common Good Approach

Core Idea:

Ethical decisions should promote values and conditions that benefit everyone in a community or society.

Key Question:

Does this action strengthen the community and promote the common good?



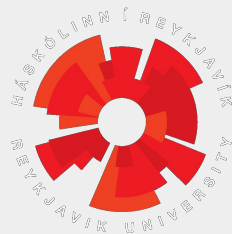
The Common Good Approach

Decision Making Focus:

- Emphasize shared values like safety, education, environment
- Supports civic responsibility and public trust

Useful when:

- An issue affects public institutions, services, or infrastructure
- You want to align actions with long-term collective well-being



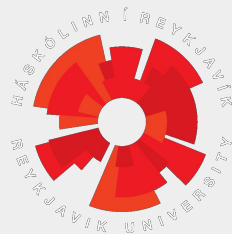
The Virtue Ethics Approach

Core Idea:

Focus not on rules or outcomes, but on the moral character of the person making the decision. What would a virtuous person do?

Key Question:

What would a person with good character do in this situation?



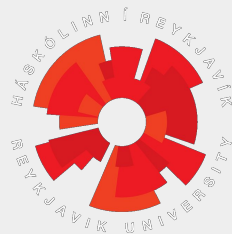
The Virtue Ethics Approach

Decision Making Focus:

- Cultivate virtues like honesty, compassion, courage, humility
- Encourages moral maturity and self-reflection

Useful when:

- There's moral ambiguity or conflicting duties
- You want to reinforce ethical leadership or professionalism



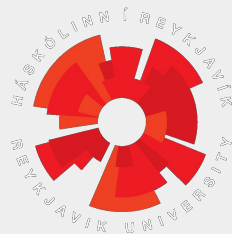
The Care Ethics Approach

Core Idea:

Ethical decision-making should emphasize relationships, empathy, and care; especially for those who are vulnerable.

Key Question:

How will this decision affect the people I am responsible for or connected to?



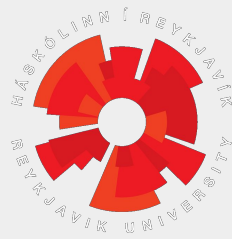
The Care Ethics Approach

Decision Making Focus:

- Prioritize context, emotional bonds, and responsibilities of care
- Recognizes the moral value of dependency, trust, and sensitivity

Useful when:

- You're dealing with unequal power dynamics
- Ethics must account for real-world human needs and emotions



The Six Lenses of Ethical Decision Making

Their Usage:

You can use each lens

- Individually: e.g., “Apply the justice lens to this AI hiring system”
- Comparatively: e.g., “Compare the rights and utilitarian responses”
- Iteratively: e.g., “Apply three lenses to see where they agree or disagree”