AI, Ethics, and Society

21.10.2026

Recap

We talked about empirical learning.

Learning based on experiences through interactions with the task-environment.

Experimentation for knowledge acquisition.

Happening under limited time and energy.

Requirement for continuous learning in a vastly complex world.

Recap

We also talked about what kind of knowledge the agent needs to extract.

Correlations help for predictions, but are insufficient for creating reliable plans.

Causation enables actions (causing effects that the system wants to see in the world).

Recap

Explanation vs Self-Explanation.

Explanations depend on causation.

The question “Why?” asks for the causes of a decision, change of the environment, or generation of output.

Correlational data is insufficient for a good explanation.

Self-Explanation is the ability to not only explain outside phenomena, but also the system’s own actions, knowledge, understanding, beliefs, reasoning, etc.

Self-Explainability subsumes Explainability.

Recap

The Explanation hypothesis (ExH) states that explanations are fundamental elements in all advanced learning.

Explanations are the stuff that the system needs to come up with in the first place.

Not an afterthought!

The plan the system should commit to is the one that represents the best explanation why it should work.

Recap

Lastly, we (that is I) talked about trustworthiness.

Trustworthiness depends on reliable predictability of what the system does.

In advanced AI, the extremely high number of unbound variables of the environments these systems can tackle, makes this a real problem.

How can we achieve trustworthiness?

Probably some sort of certification process for a particular AI in a particular context performing a particular task.

Next Generation AI

21.10.2026

Next-Generation AI

- Cognitive Autonomy
- Empirical Reasoning
- Cumulative Learning
- Cognitive Growth

Cognitive Autonomy

What is needed for cognitive autonomy?

- Selection
  - Autonomous selection of variables:
    Decide from a set of variables with potential relevance, whether one of them
    - Is relevant,
    - If so how much,
    - And in what way.
  - Autonomous selection of processes:
    Decide what kind of learning algorithms to employ (learning to learn)

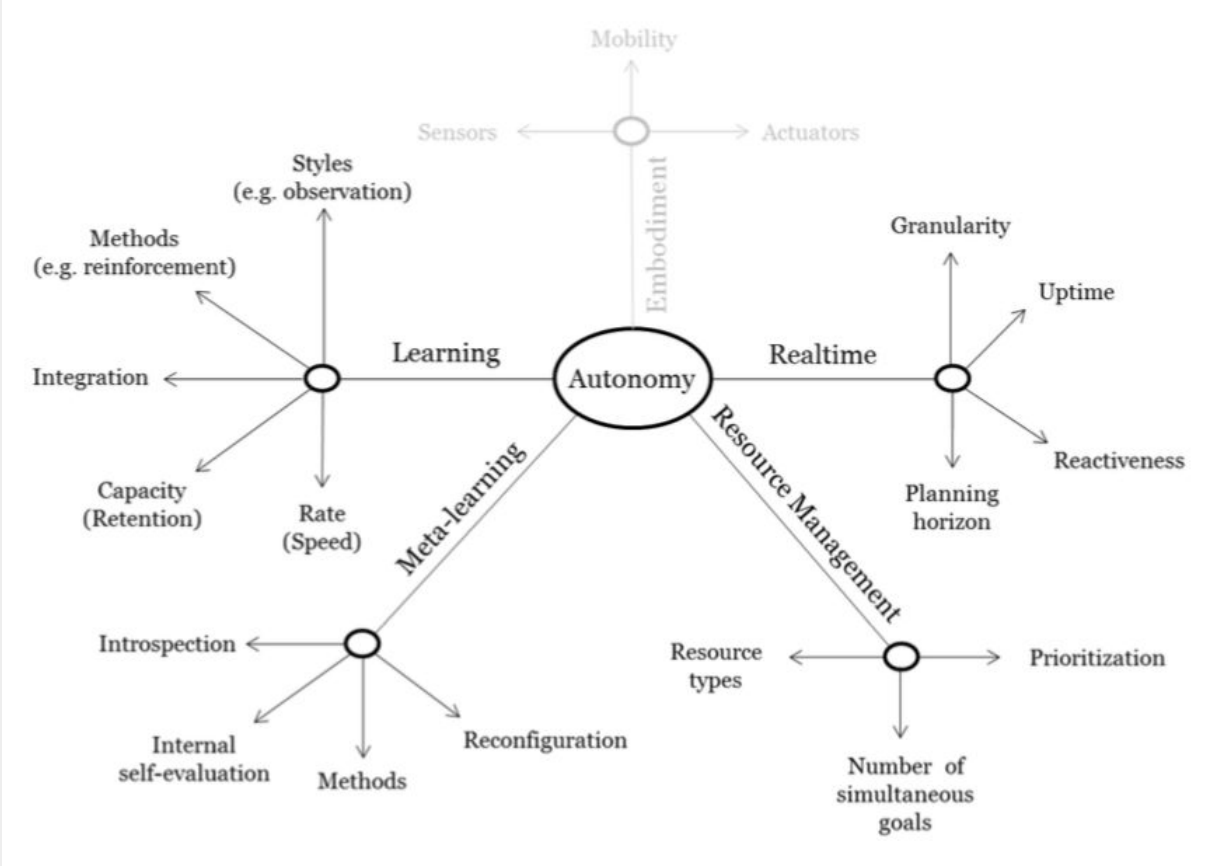
Very few (if any) learning methods exist that can do either.

Cognitive Autonomy

- Goal generation
  - Very few (if any) learning methods exist that can generate their own (sub-) goals.
  - Of those that might be said to be able to, none can do so freely for any topic or domain
- Control of resources
  - Resources are at the very least:
    - Computing power (think time)
    - Time
    - Energy
  - Few (if any) learning methods are any good at
    - Controlling their resource use
    - Planning for it
    - Assessing it, or
    - Explaining it.

Cognitive Autonomy

- Novelty
  - To handle novelty autonomously, a system needs autonomous hypothesis creation related to variables, relations, and transfer function.



"Autonomy comparison framework focusing on mental capabilities. Embodiment is not part of the present framework, but is included here for contextual completeness." From Thorisson & Helgason 2012 Source

Empirical Reasoning

Empirical Reasoning Types:

- Deduction
- Abduction
- Induction
- Analogy
- Axiomatic Reasoning
- Non-Axiomatic Reasoning

Empirical Reasoning

Deduction:

- Figuring out implication of facts (or predicting what may come)
- From general to specific
- Producing implications from premises
- The premises are given, the work involves everything else
- Conclusion is unavoidable given the premises (in a deterministic, axiomatic world)

Empirical Reasoning

Abduction

- Figuring out how things came to be the way they are.
  - Or: How particular outcomes could be made to come about
  - Or: How particular outcomes could be prevented
- The outcome is given, the work involves everything else
- E.g., Sherlock Holmes

Empirical Reasoning

Induction:

- Figuring out the general case
- From specific to general
- Making general rules from a (small) set of examples.
- E.g.: The sun has risen in the east every morning up until now, hence, the sun will also rise in the east tomorrow.

Empirical Reasoning

Analogy:

- Figuring out how things are similar or different.
- Making inferences about how something X may be (or is) similar through a comparison to something else Y.
  - Can be done on shared properties (X and Y share some properties)
  - Or through shared processes (X and Y behave similarly)

Empirical Reasoning

Axiomatic Reasoning

- When the previously stated methods are used for data and situations where all the rules and data are known and certain.
- This form of reasoning has a long history in mathematics and logic

Non-Axiomatic Reasoning

- In the physical world, the “rules” are not all known and not all certain.
- This calls for a version of the above that is defeasible.
- Any data, rule, or conclusion could be incorrect and thus defeasible.

Empirical Reasoning

Considerations:

- Why empirical?
  - The concept empirical refers to the real world: Wh live in a physical world, which is to some extent governed by rules. Some of which we know something about.
- Why reasoning?
  - Logic-governed operations are highly efficient and effective for interpreting, managing, understanding, creating and changing rules.
  - We call such operation reasoning.
  - Since we want to make machines that can operate more autonomously (e.g., in the real world), reasoning skills is one of those features that such systems should be provided with.

Empirical Reasoning

- Why empirical reasoning?
  - The physical world is uncertain because we only know part of the rules that govern it.
  - Even when having good rules (e.g., things fall down), applying such rules is a challenge.
  - Especially when faced with the passage of time.
  - The term “empirical” refers to the fact that the reasoning needed for intelligent agents in the physical world are, at all times, subject to limitations.
  - These limitations encompass at least energy, time, space, and knowledge.
  - Also called the “Assumption of Insufficient Knowledge and Resources (AIKR)” (by Pei Wang)
- Trustworthy Reasoning
  - Reasoning in the physical world requires non-axiomatic reasoning.
  - Achieving trustworthiness in non-axiomatic reasoning is a huge challenge that AI has not come to grips yet.

Cumulative Learning

Unifies several separate research tracks in a coherent form easily relatable to AI requirements:

- Multitask learning
- Online learning
- Lifelong learning
- Robust knowledge acquisition
- Transfer learning
- Few-shot learning

Cumulative Learning

Multitask learning:

- Ability to learn more than one task, either at once, or in sequence.
- The cumulative learner’s ability to generalize, investigate, and reason will affect how well it implements this ability.

- Subsumed by cumulative learning because knowledge is contextualized as it is acquired. Meaning that the system has a place and a time for every bit of information it absorbs.

Cumulative Learning

Online learning:

- Ability to learn continuously, uninterrupted, and in real-time from experience as it comes.
- Without specifically iterating over these experiences many times.

- Subsumed by cumulative learning because new information, which comes in via experiences, is integrated with prior knowledge at the time it is acquired.
- So a cumulative learner is always learning as it’s doing other things.

Cumulative Learning

Lifelong Learning

- Learning and integrating new knowledge throughout the operational lifetime.
- Learning is “always on”.
- We expect the “learning cycle” (alternating learning and non-learning periods) to be free from designer tempering or intervention at runtime.
- Provided this, the smaller those periods become, to the point of being considered virtually or completely continuous, the better the “learning always on” requirement is met.

- Subsumed by cumulative learning because continuous online learning is steady and ongoing all the time - why turn it off?

Cumulative Learning

Robust knowledge acquisition

- The antithesis of which is brittle learning, where new knowledge results in catastrophic perturbations or prior knowledge (and behaviour).

- Subsumed by cumulative learning because new information is integrated continuously online, which means the increments are frequent and small.
- Inconsistencies in the prior knowledge get exposed in the process.
- Opportunities for fixing small inconsistencies are also frequent because learning is lifelong which mean new information is highly unlikely to result in e.g. catastrophic forgetting.

Cumulative Learning

Transfer learning:

- The ability to build new knowledge on top of old, in way that the old knowledge facilitates learning the new.
- While interference/ forgetting should not occur, knowledge should still be defeasible.
- The physical world is non-axiomatic, so any knowledge could be proven incorrect in light of contradicting evidence.

- Subsumed by cumulative learning because new information is integrated with old, which may result in exposure of inconsistencies, missing data, etc.
- Which is dealt with as a natural part of the cumulative learning operations

Cumulative Learning

Few-shot learning

- Ability to learn something from very few examples or very little data.
- Common variants include
  - One-shot learning, where the learner only needs to be told (or experience) something once.
  - Zero-shot learning, where the learner has already inferred it without needing to experience or be told.

- Subsumed by cumulative learning because prior knowledge is transferable to new information.
- (Theoretically) only the delta between what is priorly learned and what is required for the new information needs to be learned.

Cognitive Growth

- Changes in the cognitive controller (the core “thinking” part) over and beyond basic learning.
- After a growth burst of this kind, the controller can learn differently/better/new things especially new categories of thing.
- In humans, cognitive growth seems to be nature’s method for ensuring safety when knowledge is extremely lacking.
- Instead of allowing a human baby to walk around and do things, nature makes human babies start with extremely primitive cognitive abilities that grow over time under guidance of caretakers.
- Piaget’s Stages of Development (YouTube video)

Ten Hypotheses of AGI

The next slides represent findings of the 2022 paper by Kristinn Thórisson and Henry Minsky.

Source:

The Future of AI Research: Ten Defeasible ‘Axioms of Intelligence’

Link:

https://proceedings.mlr.press/v192/thorisson22b/thorisson22b.pdf

H1: Unification

"*Intelligence is a systemic phenomenon, requiring the unification and coordination of many known and undiscovered information processing principles*." – (Thórisson & Minsky 2018)

- Human-level machine intelligence can not be achieved by only focusing on one aspect of cognition.
- Unification is not gluing of existing AI algorithms.
- There will be systems or architectures with intertwined parts.

H1: Unification

Potential ethical implications:

- Moral agency and rights:
  If an AI system demonstrates a sophisticated understanding of the world and autonomous cognitive capabilities, society might start to question whether it deserves certain rights or responsibilities.
  - Should a machine designed to act and behave like a human have certain rights?
- Unintended emergent behaviour:
  As intelligence is built up through intertwined parts, certain interactions may lead to unexpected or emergent behaviors that could be harmful or dangerous.

H2: Cumulative Learning

"*To achieve autonomy, and be capable of general learning, an agent in the physical world must be able to learn incrementally and modify its existing knowledge in light of new information.*" – (Thórisson & Minsky 2018)

- The integration of new pieces of knowledge into an agent's knowledge base in a systematic way, efficiently and effectively.
- Complex environments may change unexpectedly. Lifelong continual learning is required.
- The integration of new and old pieces of knowledge is key for autonomy.

H2: Cumulative Learning

Potential ethical implications:

- Cumulative learners are more vulnerable to being fed with misinformation, leading to
  - Safety issues: new wrong information could overwrite useful and valuable knowledge.
  - Ethical drift: when encountering new situations/environments with different conflicting ethical norms
- Privacy issues:
  The AIs will continue to collect data that could be sensitive.
- How to solve privacy and safety issues?

→ By being extremely goal-driven, recognize grey areas, and simulations

H3: Compositional Knowledge

"*To enable incremental buildup and modification of knowledge, cumulative (i.e. incremental, continual) learning requires compositional knowledge representation. When facing novel phenomena, such knowledge creation must be based on informed, contextual, focused, and defeasible hypothesis generation*." – (Thórisson & Minsky 2018)

- Instead of a giant, fixed ANN model, small pieces of knowledge that can be integrated are required.
- Causal models provide compositionality.
- The knowledge must be treated as falsifiable hypotheses.
- Context is learned and formulated in the knowledge representation
- Important: compositional knowledge representation allows for analogy-making.

H3: Compositional Knowledge

Potential ethical implications:

- Error propagation: Compositional knowledge increases the modularity of the knowledge but may lead to error propagation, as knowledge pieces are composed of other pieces.
- Bias amplification: when a knowledge piece is biased, the bias may be amplified through information processing mechanisms.

→ Evaluate the knowledge using reasoning: in-, de-, and abduction, and analogy.

- Compositionality may increase complexity and, therefore, decrease transparency.

→ Causality preserves transparency at every level of detail.

H4: Explanation

"*To keep track of arguments for and against knowledge hypotheses, a capacity for (self-) explanation is necessary. This explanation capacity is based on reasoning processes, including deduction, abduction, induction, and analogy*." – (Thórisson & Minsky 2018)

- Explanation based on causal knowledge requires logic with four fundamental processes.
  - Abduction: inferring causes from effects
  - Deduction: prediction effects from causes
  - Induction: generating cause-effect relations
  - Analogy: comparing cause-effect relations with situation-goal patterns
- Future AI will have to perform causal discovery/reasoning.

H4: Explanation

Potential ethical implications:

Only positives. Explanations increase safety and transparency, they can be used to reduce bias, making the system more fair, etc.

H5: Resource Control

"*To achieve existential autonomy in a world with limited resources, informed (self-)control of available resources is necessary, whether learned or pre-programmed.*" – (Thórisson & Minsky 2018)

- Physical tasks have deadlines.
- Agents cannot know all existing relations in the world.
  - Unknown unknowns facilitate learning and planning.
- Agents in the physical world must consider limitations on the use of energy.

H5: Resource Control

Potential ethical implications:

- Multi-goal creation by AI and prioritization of the AI's goals: An autonomous AI that self-manages resources could end up prioritizing its own goals and resources over its own tasks, certain human values, society's norms, etc.
  - E.g., To ensure continual functionality, an AI might use self-preservation behaviors that prioritize its survival over the needs or safety of humans.

What would be the potential solutions?

→ Goal conflict detection and resolution via reasoning. Predicting the outcome of committing to certain goals can be costly.

→ Some of the ground-truth values/drives must not be changed. Educate them properly.

H6: Transparency and Reflection

"*If one or more of the above hypotheses H2 to H4 are correct, transparent operational semantics are necessary for general learning, as well as for achieving cognitive growth. Achieving true autonomous artificial intelligence (including "human-level," "general machine intelligence," or any other reasonable conceptualization of a truly intelligent machine), is therefore highly likely to require reflection. Autonomous general learners must therefore be capable of reflection*." – (Thórisson & Minsky 2018)

- Reflection: Looking back at one's thoughts and actions in the past. Needed for two main reasons:
  - Understanding one's own thinking: thinking about the reasons of choosing specific thought processes or actions.
  - Improving how one thinks over time: The agent can grow cognitively, meaning it can get better at learning, solving problems, and adapting to new situations

H6: Transparency and Reflection

Potential ethical implications:

- Reflection leads to changing the reasoning processes, and making understanding the system's decisions more challenging

→ Causal reasoning AND meta-reasoning.

- Reflection leads to more self-reliance and thus more autonomy
- Thus, starting to prioritize its own goals based on its understanding of what improves itself, potentially creating goals that might diverge from human intentions, expectations and values, interests, or safety.

→ Transparency

H7: Abstraction

"*Knowledge abstraction is a key feature of general autonomous cognition—without it, handling complex information at multiple levels of detail becomes intractable.*" – (Thórisson & Minsky 2018)

- Abstraction: A cognitive process of pruning out unimportant knowledge from knowledge structures and coming up with new heuristics for solving tasks.
- Conceptual creativity: new concepts will be the result of an abstraction process
- Self-generated heuristics: Heuristics must not be given to AI systems as initial seed knowledge
- Emergent concept creation: The ability to generate new concepts arises naturally from fundamental mechanisms.

H7: Abstraction

Potential ethical implications:

- Transparency issue: Abstraction may lead to creating new concepts that are not easily interpretable to humans.

→ Teaching it natural languages.

- Biases in abstraction: abstraction involves classification, which may shift due to being in environments with values other than the society norms, leading to biases in the way the classification occurs.

→ Instructions from human teachers, as it is not safe to try all classifications in practice.

H8: Seed

"*Existential autonomic learning means freedom from a teacher, which means that the learner must be provided with a program (a 'seed'), up-front (at 'birth'), that contains actionable principles for bootstrapping its learning, and the development and growth of the cognitive apparatus over time. The more general this knowledge, the more flexible can the learning get, albeit at the cost of taking more time.*" – (Thórisson & Minsky 2018)

- Self-programming for cognitive growth: AI agents must change their programming paradigms based on their seed knowledge.
  - When the seed knowledge is too general
  - When the learned knowledge is not useful anymore.

H8: Seed

Potential ethical implications:

- If the top-level goals are defined against society's values and norms, the cumulative learning system with human-level cognitive capabilities could be extremely dangerous.

Mandate that developers implement ethical alignment protocols

H9: Holistic Approach

"*A holistic stance and approach to the phenomenon of intelligence– that is, a research program that does not dismember the topic under study, transforming it in the process to a mere facsimile of itself– is by far more likely than other methodologies to deepen our understanding of intelligence and enable the creation of truly intelligent machines.*" – (Thórisson & Minsky 2018)

- Risk of Misinterpretation: Focusing on isolated parts risks breaking important connections, and potentially altering the AI system away from its original intended form.
- In holistic approaches, cognitive functions depend on each other.

H10: Constructivism

"*A constructivist view, with an emphasis on self-guidance and self-originated meaning, is a useful methodological stance (and one of possibly only very few) to help focus on the issue of holistic systems and unify all of the above principles in a single system.*" – (Thórisson & Minsky 2018)

Constructivism is an approach in AI that relies on

- Causal representation of knowledge and reasoning
- Agency, feedback loops and control
- Compositional defeasible knowledge creation and integration
- Reasoning under insufficient knowledge and resources.

Future A(G)I Systems and the EU AI Act

Can European Union’s AI Act deal with future AI?

- The way risks are measured should be improved. Proper definitions of AGI systems should be provided. More or different categories of risks are needed.
- Dynamic safety and compliance monitoring measures are required.
- The long-term ethical implications of future AI should be scrutinized.
- More systematic education for AGI developers is required.
- Adaptation to adapting systems is required (see documentation of model parameters, etc.)