AI, Ethics, and Society

27.10.2025

Requirements for Next-Gen AI

1. Autonomy
2. Cause-Effect Knowledge
3. Cumulative Learning
4. Empirical Reasoning
5. Trustworthiness

Autonomy

- The ability of a system to “act on its own”
- Self-Inspection:
  - Being able to inspect (measure, quantify, compare, track, make use of) it’s own development for use in its continued growth - whether learning, goal-generation, selection of variables, resource usage, or other self-...
  - Virtually no systems exist of yet capable of such self-inspection
- Self-Growth:
  - Necessary for autonomous learning in task-environments with far higher complexities than the controller operating in them.
  - Even more important when certain bootstrapping thresholds are necessary before transitioning into more powerful/ different learning schemas.
  - I.e., only few bits of knowledge can be programmed in, to ensure maximal flexibility - Thus, something that protects the controller while it develops more sophisticated learning.

Autonomy

- Self-Inspection and Self-Growth are key features of autonomy that any human-level intelligence probably must have
  - Probably because 1) we don’t have any, yet, and 2) we don’t even have a proper definition of intelligence.
- Autonomous Learning:
  - There exist machines that learn autonomously (see RL, self-supervised learning, etc.)
  - But: Most of the existing ones are limited in that they
    a) rely heavily on quality selection of learning material/ environments
    b) require careful setup of training, and
    c) need careful and detailed specification of how progress is evaluated

Three Levels of Autonomy

- Level 1 - Automation
- Level 1.5 - Reinforcement Learning
- Level 2 - Cognitive
- Level 3 - Biological

Source: Thorisson 2020 - http://alumni.media.mit.edu/~kris/ftp/Seed-Programmed-General-Learning-Thorisson-PMLR-2020.pdf

Three Levels of Autonomy

Level 1 - Automation:

- Description
  - The lowest level may be called “mechanical”
- Uniqueness
  - Fixed architecture. Baked-in goals. Does its job. Does not create.
- Examples
  - Examples: Thermostats, DNNs
- Learning
  - No “ learning” after it leaves the lab.

Three Levels of Autonomy

Level 1.5 - Reinforcement Learning

- Description
  - Can change their function at runtime.
  - Cannot accepts goal descriptions.
  - Cannot handle unspecified variables
  - Cannot create sub-goals autonomously
- Uniqueness
  - “Learns” through piecewise “boolean” (good/bad) feedback
- Examples
  - Q-Learning, other RLs
- Learning
  - Limited to a handful of predefined variables

Three Levels of Autonomy

Level 2 - Cognitive

- Description
  - Handling of novelty.
  - Figures things out.
  - Accepts goal descriptions.
  - Generates goal descriptions
  - Creates
- Uniqueness
  - Flexible representation of self
  - High degree of self-modification
- Examples
  - Humans, Parrots, Dogs
- Learning
  - Learns after it leaves the lab.

Three Levels of Autonomy

Level 3 - Biological

- Description
  - Adapts.
- Uniqueness
  - Is alive.
  - Sobject to evolution.
  - Necessary precursors to lower levels.
- Examples
  - Living creatures
- Learning
  - Adapts after it leaves the primordial soup.

Empirical Learning

When information comes from measurements in the physical world it is “empirical evidence”.

Empirical Learning is this learning based on empirical data.

- Experience-Based Learning
- Experimentation
- The Physical World
- Limited Time and Energy
- Why it matters.

Empirical Learning

Experience-Based Learning

- Learning is the acquisition of knowledge for particular purpose.
- When this acquisition happens via interaction with an environment it is experience-based

Experimentation

- To produce data needed for learning about an environment not fully known a-priori.
- Environment may prevent knowledge of all its “rules” like the physical world.

Empirical Learning

The Physical World

- The world we live in (“real world”)
- Highly complex.
- Rarely, if ever, we have a perfect model of how it behaves when we interact with it (whether it is to experiment or simply achieve some goal like buying bread).

Empirical Learning

Limited Time and Energy

- Any agent faces a major limitation when modeling the real world: the state space is enormous..
- It far exceeds any known agent’s memory capacity, even for relatively simple environments.
- Precomputing all possible outcomes or model details is impossible due to memory constraints.
- Even if enough memory were available to store all precomputed information, the retrieval of relevant data in real time would become a bottleneck.
- As the state space grows, the demands on retrieval speed increase dramatically, making timely access impractical.

Empirical Learning

Why it matters:

- Under Limited Time and Energy (LTE) in a plentiful environment, it is impossible to know all at once, including causal relation.
- Therefore, most of the time, an intelligent agent capable of some reasoning, will be working with uncertain assumptions where nothing is certain.
- Only some things are more probable than others

Correlation, Causation, and AI

- Correlation
- Causation
- Causal Models
- State of the Art

Correlation, Causation, and AI

Correlation

- The apparent relationship between two or more variables.
- When observed repeatedly, the value of one seems to follow the other (and vice versa)
- Correlation is not directional (we don’t know which one causes the other)
- Sufficient for simple prediction
    - if A and B correlate highly, then it does not matter if we see an A or a B, we can predict that the other is likely on the scene.

Correlation, Causation, and AI

Causation

- Directed relationship between two (or more variables) A and B.
- Such that if you change the value of variable A, then the value of variable B changes also, according to some function.
- NOT (necessarily) vice versa.
- Supports action
  - If we know A causes B and we want to see B disappear, and know how to make A disappear, we can act and make A disappear.
- Subsumes correlation.

Correlation, Causation, and AI

Causal Models

- Necessary to guide action
- While correlation might give us an indication of causation, the direction of the “causal arrow” is critically necessary for guiding action.
- Knowing which way the arrow points is usually not too hard to find out through empirical experimentation

Correlation, Causation, and AI

State of the Art

- Judea Pearl:
  - Most fervent advocate of causality in AI, and the inventor of the Do Calculus.
- Recent work by Judea Pearl demonstrates clearly the fallaciousness of the statistical stance.
- Fixes some important gaps in our knowledge on this subject.
- Hopefully will rectify the situations in the upcoming years.

YouTube lecture by Judea Pearl on causation:

https://www.youtube.com/watch?v=8nHVUFqI0zk

Explanation and Explainability

The ability to explain, after the fact, during, or before, why something happened the way it did, how it could have happened differently but didn’t, and why.

In AI:

- The ability of a controller to explain, after the fact, during, or before, whit it did something ir intends to do it.

Explanation depends on causation!

It is impossible to explain anything in any useful way without referring to general causal relations.

Explanation and Explainability

If a controller does something we don’t want it to repeat - e.g. crash a car - it needs to be able to explain why it did what it did.

If it can’t, it mean it - and thus, we - can never be sure of why it did what it did, whether it had any other choice, or under which conditions it might do it again.

Again: No explanation without causation.

Discernible causal structure is a prerequisite for the kind of explainability we aim for.

Explanation and Explainability

Bottom Line for human-level AI:

- To grow, learn and self-inspect, an AI must be able to sort out causal chains.
- If it can’t it will be incapable not only to explain to others why it is as it is, but also to itself why things are the way they are.
- Thus, it will be incapable of sorting out whether something it did is better for its own growth than something else.
- Explanation is the big black-hole of ANNs. They are black-boxes and thus, in principle unexplainable - whether to others or themselves.
- One way to address this is by encapsulating knowledge as hierarchical models that are built up over time and can be deconstructed when needed.

Self-Explaining Systems

Explainability != Self-explainability

- If an intelligence X can explain a phenomenon Y, Y is explainable by X through some process chosen by X.
- In contrast, if an intelligence can explain itself, its own actions, knowledge, understanding, beliefs, and reasoning, it is capable of self-explanation.
- The latter is stronger and subsumes the former.

Self-Explaining Systems

Why this matter more than you might think:

- The Explanation Hypothesis (ExH) states that explanation is in fact a fundamental element in all advanced learning.
- Explanation is a way to weed out alternative (and incorrect) hypotheses about how the world works.
- For instance, if the knowledge already exists in a controller to “do the right thing” - that is, for the right reason - in an emergency situation, the explanation of why it does what it does already exists embedded in its knowledge!

See Thórisson 2022 - https://proceedings.mlr.press/v159/thorisson22b/thorisson22b.pdf

Trustworthiness

The ability of a machine's owner to trust that the machine will do what it is supposed to do.

- Any machine created by humans is created for a purpose.
- The more reliably it does its job (and nothing else) - and does it well, the more trustworthy it is.
- Trusting simple machines (like thermostats) involves mostly durability since they have very few open variables (at time of manufacturing). Their task is well defined and well known and their reasonably precise operation can be ensured with simple engineering.

Trustworthiness

In AI:

- AI is supposed to handle diversity in one or more tasks.
- A learning AI system leaves the machine's task undefined at manufacturing.
- The smarter an AI system, the more diversity it can handle.
- A requirement should be that "trustworthiness grows with the mindpower of the machine"

Trustworthiness

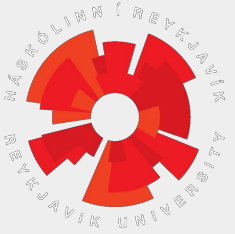
In human-level AI:

- Extremely high number of unbound variables at manufacturing time.
- What does trustworthiness mean in this context?
- We can look at human trustworthiness: Numerous methods exist for ensuring trustworthiness.
  - E.g., license to drive, air traffic controller training, certification programs, etc.
- These exist for all humans because their principles of operation are shared at multiple levels of detail (biology, sociology, psychology).
- For an AI, this is different because the variability in the makeup of the machines is enormous.

Trustworthiness

Achieving trustworthiness (in human-level AI):

- Requires reliability and predictability at multiple levels of operation.
- Trustworthiness can be ascertained through special certification programs geared directly at the kind of robot/ AI system in question.
- Kind of like certifying a particular horse as safe for a particular circumstance and purpose, e.g., horseback riding kids.



Questions?