Philosophy and AI
Lecture 4: Epistemology - Kant, Stability, Formal Learning Theory

Marco Degano

12 February 2025

Readings

Required:

- Glymour 2015: chapter 9, pp. 211 – 219; 227–229
  - Required: Introduction; The Kantian Picture; Constructional Systems; The History of Knowledge
  - Optional: Conventionalism and Analytic Truth; Doubts.
- Glymour 2015: chapter 10, pp. 239 – 243; 250 – 254; 259
  - Required: Introduction; Knowledge; Putnam's Framework; Conclusion.
  - Optional: Reliability and Justification; The Mathematics of Reliability.

Optional:

- Schulte, Oliver, "Formal Learning Theory", The Stanford Encyclopedia of Philosophy `https://plato.stanford.edu/archives/sum2024/entries/learning-formal/`
- Video clip on tracking theory by *Wireless Philosophy* `https://youtu.be/lyE0xiMjaoI?si=bIdQ6dY4Lrl3wsqm`

Outline

1. Kant
2. Stability
3. Formal Learning Theory
4. Summary

Introduction

In the last lecture, we saw the Bayesian solution to the problem of induction:

- replace absolute knowledge by degrees of rational belief and
- focus on how we can, though observation, reliably and quantifiably convergence to the truth.

Introduction

In this lecture, we consider three other solutions:

(1) Kant: We can observe the world only through our cognitive system, and this is such that it supports inductive inference.

(2) Stability: A crucial feature that turns true belief into knowledge is their stability, i.e., remaining true beliefs in similar situations.

(3) Learning: If we precisely analyze the process of testing hypotheses based on observations, we find sensible notions of reliable confirmation.

Outline

1. Kant
2. Stability
3. Formal Learning Theory
4. Summary

Kant's epistemology in a nutshell

- **The Limits of Perception:** We do **not** observe reality **directly**; rather, we perceive it through the structures imposed by our cognitive faculties.
    - External objects generate sensory input (e.g., light waves enter the eye and trigger electrical signals in the retina).
    - The brain processes this raw, fragmented, and often ambiguous sense data, constructing a coherent representation of the world.
    - This interpretation happens *automatically* and is largely beyond our conscious control.
    - Kant refers to raw sensory input as the manifold and the process of organizing it into meaningful experience as synthesis.

Kant's epistemology in a nutshell

- **The Constructed Nature of Experience:** The world we experience is *not* a direct reflection of objective reality but rather a construction of our cognitive faculties.
    - Certain *structural* features of experience (such as time, space, and causality) are not derived from observation but are *built into* our cognitive system.
    - Since our perception is structured this way, all our experiences need to conform to these structures.
    - Kant argues that these fundamental organizing principles make empirical knowledge possible.

Kant's epistemology in a nutshell

- **Kant's Transcendental Argument:** Why must experience necessarily conform to these structures?
  - Example: Space is necessarily three-dimensional with Euclidean geometry.
  - No experience can contradict this because spatiality is part of how we organize perception.
  - Kant describes such organizing principles as necessary conditions for possible experience.
  - These are known as synthetic a priori truths:
    - *Synthetic* (not derivable purely from logic or meaning).
    - *A priori* (known independently of experience).

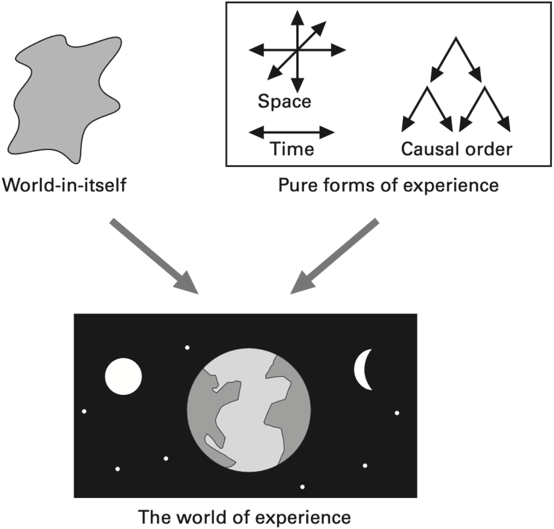

Figure 9.1
Kant's picture of metaphysics.

Interlude: Analytic vs. Synthetic & A Priori vs. A Posteriori

- Analytic a priori:
  - Judgments whose truth is determined by the meanings of the terms and logical form.
  - Known independently of experience.
  - **Example:** "All bachelors are unmarried men."
- Synthetic a priori:
  - Judgments that are necessarily true and known independent of experience, yet their truth is not contained in the meanings of the terms.
  - **Example:** "$7 + 5 = 12$"
- **Analytic judgments** merely unpack the content already contained in the concepts (e.g., definitions).
- **Synthetic judgments** extend our knowledge by connecting concepts in new ways.

Interlude: Analytic vs. Synthetic & A Priori vs. A Posteriori

A priori knowledge is independent of experience, while **a posteriori knowledge** requires empirical evidence.

- Synthetic a posteriori:
  - Judgments whose truth depends on empirical observation and where the predicate adds new information to the subject.
  - **Example:** “The cat is on the mat”
- Analytic a posteriori:
  - According to Kant, this category is empty; all analytic judgments are known a priori.

Kant's epistemology in a nutshell

- This is how Kant replied to Hume's inductive skepticism:
  - Kant gave a transcendental argument for causality: i.e., that our cognitive system organizes all experience in a causal order.
  - More precisely, by virtue of how our cognitive system operates, every experienced event has a cause, and experienced causal connections can be expressed as a general law (e.g., heavy objects fall to the ground).
  - Being a transcendental argument, this is not a 'logical'/deductive argument for induction (as Phyrrhonians argued is impossible).
  - And it also is not circular (as Hume argued is impossible) because the justification for the inductive inference is not circularly given by the usual reliability of experience.

Kant's epistemology in a nutshell

- However, Kant still kept a form of skepticism:
  - We can never access the world in itself (*noumena*)-only the world as it appears to us (*phenomena*).
  - This means we cannot know whether our mental representations *accurately* correspond to mind-independent reality.
  - Example: We infer external objects from sense data, but we cannot know with certainty what these objects are *in themselves*.
  - This leads to a form of transcendental idealism-a middle ground between realism and radical skepticism.

Kant's influence: On AI

- Philosophers and logicians, including Bertrand Russell, Ludwig Wittgenstein, and Rudolf Carnap, sought to clarify Kant's synthesis process in formal terms.
- Russell, for example, developed a *logical language* where variables (x) represent raw sense data and predicates ($R(x)$) describe higher-level concepts (e.g., *redness*, *tree-ness*).
- Carnap spelled these ideas out in 'The logical construction of the world' (1929), giving a computational procedure to synthesize higher-level concepts from lower-level ones.
- Carnap spoke of constructional systems. (Some translate it as constitutional systems.)
- Carnap's ideas resemble modern AI techniques that layer abstraction, progressively transforming raw data into meaningful structured knowledge.

Kant's influence: On AI

- This idea of describing cognition (or processes more generally) with logical tools is also driving symbolic AI and motivating programming languages like PROLOG.
- But also, for example, DeepMind research scientist Richard Evans used Kant's ideas on synthesis to restrict the search space sufficiently to make it possible to find logical theories explaining observed raw data (*The Apperception Engine*).

  Kim, H. & Schönecker, D. (2022). *Kant and Artificial Intelligence*. `https://doi.org/10.1515/9783110706611`

Carnap's Ideas in Machine Learning: Feature Abstraction

- **From Raw Data to Meaningful Representations:** In machine learning, high-dimensional input data undergoes processing where patterns and relationships are extracted, much like how Carnap envisioned conceptual construction.
- **Feature Engineering as Synthesis:**
  - Machine learning models rely on feature extraction to transform raw input into structured, meaningful variables.
  - This mirrors Carnap's constructional systems, where primitive observations (e.g., pixel values, audio signals) are synthesized into higher-level concepts (e.g., objects, speech patterns).

Carnap’s Ideas in Machine Learning: Feature Abstraction

- **Neural Networks and Layered Abstraction:**
  - Deep neural networks progressively abstract features from data across multiple layers.
  - Early layers extract basic patterns (e.g., edges in images, phonemes in speech).
  - Deeper layers construct hierarchical representations (e.g., recognizing faces, understanding sentence structure).
  - This process reflects Kantian synthesis, where cognition structures sensory input into complex perceptions.

Carnap's Ideas in Machine Learning: Feature Abstraction

- **Dimensionality Reduction and Conceptual Synthesis:**
    - Techniques like Principal Component Analysis (PCA) and autoencoders reduce data complexity while preserving essential structure.
    - This resembles Carnap's notion of organizing knowledge into an efficient, structured hierarchy.
    - Example: PCA can compress thousands of variables into a smaller set of meaningful dimensions, akin to how our minds extract core concepts from varied experiences.

Carnap's Ideas in Machine Learning: Feature Abstraction

- **Applications of Carnap's Insights in AI:**
  - Image Recognition: CNNs detect edges, textures, and eventually entire objects, following a structured synthesis of visual data.

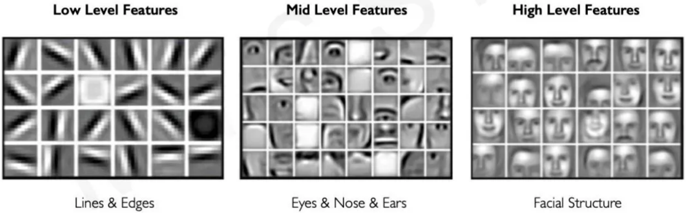

Image source

  - Explainable AI (XAI): Understanding how features contribute to predictions aligns with Carnap's goal of making conceptual synthesis explicit.

Kant's Influence: Conventionalism

The Challenge to Kant's Necessity of Cognition:
Kant argued that certain fundamental structures of cognition (e.g., space, time, causality) are necessarily imposed by the mind. However, later thinkers questioned this.

Kant's Influence: Conventionalism

- **The Rise of Conventionalism:**
  - Conventionalists, such as Henri Poincaré and Hans Reichenbach, argued that these cognitive structures are not *necessary*, but rather conventions we adopt for practical or theoretical reasons.
  - **Example:** Instead of assuming that space is necessarily Euclidean (as Kant did), we can choose to interpret physical observations through a non-Euclidean framework.
  - If empirical observations seem to contradict Euclidean geometry, we may revise our physical laws rather than abandon Euclidean space itself.
  - This reframes Kant's synthetic a priori truths-such as the structure of space and causality-not as necessary conditions for experience but as human conventions in organizing knowledge.

Kant's Influence: Conventionalism

- **Kuhn and the Problem of Incommensurability:** Thomas Kuhn, in *The Structure of Scientific Revolutions* (1962), takes one step further and argues that different scientific paradigms are incommensurable:
    - Competing scientific frameworks (e.g., Newtonian mechanics vs. Einsteinian relativity) rely on different conceptual structures, making direct comparison difficult.
    - Each paradigm comes with its own methods, standards, and language, making transitions between them revolutionary rather than incremental.
    - This challenges the Kantian idea that all rational agents must perceive the world in the same fundamental way.

Kant's Influence: Conventionalism

- **Conceptual Relativism and the Loss of Universality:** Some philosophers extend Kuhn's ideas into conceptual relativism, arguing that:
  - Different individuals or cultures might have fundamentally different conceptual schemes (system of concepts which shape or organize our experience).
  - If true, this undermines Kant's claim that human cognition is universally structured in a single way.
  - It also raises questions for AI: If artificial intelligence systems develop their own conceptual schemes, can they ever fully "understand" human cognition?

Outline

1. Kant
2. Stability
3. Formal Learning Theory
4. Summary

Stability and Tracking

- Moving beyond Kant, we now examine an alternative response to skepticism: stability and tracking theories of knowledge.
- **Can We Refine the Justified True Belief (JTB) Analysis?** Skeptical challenges suggest that having a justified true belief (JTB) is not sufficient for knowledge. What additional conditions might be required?

Stability and Tracking

- **The Role of Stability in Knowledge:** Socrates, as reported by Plato, argued that knowledge must be tied down to prevent it from drifting away:
  "Once [true opinions] are tied down, they become knowledge and are stable. That is why knowledge is something more valuable than right opinion."

  (Glymour 2015, p. 244)
- This idea has influenced modern epistemology, leading to the safety condition for knowledge (Williamson, Sosa, and others):
  - If A knows that p, then A could not have easily believed p if it were false.
  - In other words, knowledge must be resistant to error in nearby possible situations.

Stability and Tracking

- **The Problem of Gettier Cases:** The safety condition helps explain why Gettier cases fail as knowledge:
  - Example (recall): Suppose you glance at a stopped clock at exactly noon. Your belief that "it is noon" is true, but only by coincidence.
  - Since the clock was not actually functioning, your belief is not stable- had you looked at a different time, you would have been wrong.
- Thus, knowledge requires not just truth and justification, but also stability against accidental correctness.

Stability and Tracking

- **Nozick's Truth-Tracking Theory:** Similarly, Robert Nozick proposed an account of knowledge that builds on stability:
  - A knows that p if and only if:
    1. p is true.
    2. A believes p via a method M.
    3. If p were false, M would not lead A to believe p. (Sensitivity condition)
    4. If p were true, M would lead A to believe p. (Adherence condition)
  - These conditions ensure that belief tracks the truth across different possible worlds.

Stability and Tracking

- **Understanding Truth-Tracking Through Counterfactuals:**
  - Statements like “If A weren’t the case, then B would have been the case” are known as counterfactuals.
  - Nozick’s theory is a counterfactual analysis of knowledge: knowledge must hold up under hypothetical variations of reality.
- **Examples:**
  - **Gettier cases:** If you had looked at the stopped clock a minute earlier, your belief would have been false, failing the sensitivity condition.
  - **Everyday knowledge** like “I have hands”: If $p =$ “I have hands” then I know p because in nearby possible worlds, my method of checking (looking at my hands) would not make me believe that I have hands.

Stability and Tracking

But:

- $q =$ "I am not a brain in a vat."
- If I were a brain in a vat, my perceptual methods would still lead me to believe that I am not a brain in a vat.
- This means my belief in q is not sensitive to its falsity, so under Nozick's theory, I do not know that I am not a brain in a vat.

Stability and Tracking

- **A Major Objection: The Closure Problem** Nozick's account violates a widely accepted principle:
    - Closure Under Known Implication: If I know p, and I know that p implies q, then I should also know q.
    - However, under Nozick's framework:
        - I know that "I have hands" (p).
        - I know that "If I have hands, then I am not a brain in a vat" ($p \to q$).
        - Yet I do not know that "I am not a brain in a vat" (q).

Outline

1. Kant
2. Stability
3. Formal Learning Theory
4. Summary

Formal Learning Theory: A Response to Skepticism

- **A Third Response to Skepticism:** We now explore formal learning theory, a framework for analyzing how we form and test hypotheses based on observations.
- **Key Question:** Which hypotheses, if true, can be learned or confirmed, and in what precise sense?
- To illustrate, let's consider a concrete example.

Formal Learning Theory: A Response to Skepticism

- **Example: Learning the Hypothesis "All Ravens Are Black"**
  - Suppose we want to determine the truth or falsity of the universal hypothesis:

$$H = \text{“All ravens are black.”}$$

  - We observe objects in the world, seeing whether they are ravens and, if so, whether they are black.
  - We will eventually observe all objects, but possibly with repetitions.
  - At each step, we must make a conjecture about whether H is true or false, based on our observations so far.
- **Key Question:** What rule or strategy should we use to decide whether H is true?

Formal Learning Theory: A Response to Skepticism

- **Defining a Learning Rule:**
    - A rule R is a function that takes a finite sequence of observations and outputs either:
        - 1 (true) if it believes H is true, or
        - 0 (false) if it believes H is false.
    - We want R to be reliable.
- **Putnam's Criterion for Reliable Learning:** A rule R is reliable if, regardless of the order in which we observe objects:
    - If H is true, then after a finite number of observations, R will only output 1.
    - If H is false, then after a finite number of observations, R will only output 0.
- **Key Insight:** After a finite number of mistakes, we eventually converge to the correct answer and never change it again.

Formal Learning Theory: A Response to Skepticism

- **Example of a Learning Rule: Karl Popper's Approach**
  - Popper suggested a simple strategy:
    - Start by assuming H is true.
    - If an observation contradicts H (i.e., a non-black raven appears), immediately conclude that H is false and never change back.
  - This rule is reliable because:
    - If H is true, it always outputs 1 and never makes a mistake.
    - If H is false, it will eventually observe a counterexample and permanently switch to 0.
  - However, this is not the most optimal rule-better rules exist with even stronger reliability guarantees.

Formal Learning Theory: A Response to Skepticism

- **Challenging Inductive Skepticism:**
  - Formal learning theory provides a rigorous way to justify inductive reasoning.
  - Unlike traditional inductive skepticism (which argues that no amount of observation can fully confirm a universal claim), Putnam's approach shows that:
    - We may never know for certain when we have the right answer.
    - However, we can prove that a reliable learning rule will eventually settle on the truth.

Formal Learning Theory: A Response to Skepticism

- **Logical Characterization of Learnable Sentences:**
  - Putnam proved that learnable hypotheses correspond to a special class of logical sentences:
    - $\forall x \exists y \ \varphi(x,y)$
    - $\exists x \forall y \ \varphi(x,y)$

    where φ is a quantifier-free formula.
  - In mathematical logic, these are known as Δ_2^0 formulas in the arithmetical hierarchy.
  - Putnam called them trial-and-error predicates-they allow us to refine our beliefs based on accumulated evidence.

Formal Learning Theory: A Response to Skepticism

- **Beyond Putnam: The Growth of Formal Learning Theory**
  - Since Putnam's work, formal learning theory has expanded into a broader mathematical discipline.
  - Modern developments include:
    - Computational learning theory (e.g., Probably Approximately Correct (PAC) learning).
    - Algorithmic approaches to scientific discovery.

Outline

1. Kant
2. Stability
3. Formal Learning Theory
4. Summary

Summary: Three Responses to Skepticism

- We have explored three distinct responses to skepticism about knowledge. Each of these approaches, in some way, relaxes the strict reliability requirement for knowledge acquisition, allowing us to justify belief in different ways.

Summary: Three Responses to Skepticism

- **1. Kantian Epistemology: Knowledge as a Cognitive Framework**
    - Kant argues that we do not perceive reality *directly*; instead, our cognitive faculties structure experience.
    - This response sidesteps skepticism by claiming that certain fundamental structures (e.g., space, time, and causality) are necessary for experience itself.

Summary: Three Responses to Skepticism

- **2. Stability and Tracking: Knowledge as Robust True Belief**
  - Stability-based theories (such as Nozick's truth-tracking approach) argue that knowledge must be resilient to error.
  - Instead of requiring absolute certainty, these theories only demand that beliefs remain true across *similar* possible worlds.
  - However, the challenge of closure (i.e., inferring knowledge of logical consequences) remains an issue.

Summary: Three Responses to Skepticism

- **3. Formal Learning Theory: Knowledge as Reliable Convergence**
  - Instead of seeking immediate certainty, formal learning theory allows us to revise our beliefs over time.
  - We do not need to know when we have reached the truth-only that our learning process is guaranteed to eventually stabilize on the correct answer.
  - This approach provides a mathematical foundation for inductive reasoning, showing how we can learn universal truths through repeated observation and refinement.
  - However, it does not eliminate all uncertainty-there is always a period of time before convergence.

Final Reflections: From Epistemology to Modern AI

- **Emerging Challenges and Open Questions:** The intersection of philosophy and AI prompts several key questions:
  - Can AI systems truly "know" in the philosophical sense, or are they merely simulating epistemic processes?
  - As AI systems become more autonomous, will they develop their own conceptual schemes, akin to the paradigm shifts discussed by Kuhn?

Exercises

1. Sometimes we change our mind about how to make sense of sense data: e.g., what we thought was one big objects really was two; what we thought was a cause actually wasn't. Is this a problem for Kant, or can you think of a way to accommodate this in his theory? (Hint: does synthesis have to be deterministic?)
2. Why does Nozick's analysis only work for empirical knowledge and not for mathematical knowledge?
3. What do you think of the objection to Nozick's analysis: Is the closure under known implication really something we should have?
4. Discuss Putnam's criterion for reliable confirmation: how good is it if we cannot know when we're done changing our mind?