Grounding knowledge in subjective experience

Rich Sutton
University of Alberta
Is the ultimate meaning of a representation what it means to the agent? or what it means to its human designers?

Reinforcement Learning
subjective, private representations

Knowledge Representation
objective, public representations
The Problem

- How can we represent complex, commonsense knowledge of the world?

- With mathematical clarity
  - With meaning is as clear as that of a transition probability

- In such a way that it is maintainable without continuous human intervention

- In such a way that it can be learned and used flexibly (e.g., for planning)
The key to a successful AI is that it can tell for itself whether it is working correctly.
Experience (the data of AI)

The temporal stream of lowest-level sensori-motor experience

\( \Delta t = 50\text{ms} \)
Experience matters

- Experience is the most prominent feature of the computational problem we call AI.
- It’s the central data structure.
- It has a definite temporal structure.
  - revealed and chosen over time
  - speed of decision is important
  - order is important
- This has unavoidable implications for AI.
Experiential knowledge hypothesis:

All world knowledge is a prediction or memory of sensori-motor experience

- Knowledge is subjective
- Knowledge is ultimately low-level
- Logic and math are not world knowledge
  - they are true in any world
A Grand Challenge: Grounding knowledge in experience

- To represent human-level world knowledge solely in terms of lowest-level experience
  - sensations
  - actions
  - time
- A minimal ontology
  - no objects, no people, no space, no self, no chickens...
  - all these are “just” patterns in sensation & action
What would it be like to accept the challenge?

- Abstraction is key
  - abstract states (e.g., predictive representations)
  - abstract actions/transitions (e.g., options)
- Need to think in unfamiliar ways
- Microworlds, robotics
- Indexical (deictic) representations
  - sequence instead of symbols
In subjective terms,

- What is space?
  - regularities in sensation change with eye movement
- What are objects?
  - subsets of sensations
  - that tend to occur together temporally
  - and can be in arbitrary relative spatial arrangements
• What is my body, my hands?
  - objects that are always present
  - and can be controlled

• What are people?
  - objects that may move on their own
  - that have a particular subset of sensations
  - whose presence may change my sensations for the better
  - eventually:
    ✦ that are best predicted with respect to goals
    ✦ that are analogous to me
What would it be like to accept the challenge?

- Abstraction is key
  - state
  - time/transitions
- Need to think in unfamiliar ways
- Microworlds, robotics
- Indexical (deictic) representations
  - sequence instead of symbols
Relational ⇒ Indexical

∀ objects X, If I drop X, then X will be on the floor

- Holding object X means predicting certain sensations if, for example, one directs one’s eyes toward one’s hand
- Thus, on dropping, the predicted sensations are merely transferred from the looking-at-hand prediction to the looking-at-floor prediction
- Such transfer of existing predictions should be a common part of visual knowledge - updated every time the eyes move

∃ X, Y, such that Red(X), Blue(Y), and Above(X, Y)

- There is some place I can foveate and see Red
- There is some place I can foveate and see Blue
- If I foveate first the Red place, “mark” it, then the Blue place, the mark will be above the fovea (repeat until succeeds)

These are typical ideas of modern, active, deictic vision
Explicit Prediction Manifesto

Every prediction is a question and an answer, and both the question and the answer must be *explicit* in the sense of being accessible to the AI agent, i.e., of being machine readable, interpretable, and usable.
Temporal-difference networks

- Main idea: separate the problem of prediction into questions and answers, two networks

- The question network represents the explicit meaning of predictions
  - inter-predictive temporal relationships
  - can be used to represent a wide range of compositional, abstract, predictive questions

- The answer network computes estimates of the predictions
sensation: color ahead
actions: L(eft), R(right), F(orward)
options: Leap (to wall), Wander (randomly)
Pros and cons of subjective grounding of knowledge

- **Loses**
  - easy expressiveness
  - coherence with people
  - interpretability, explainability

- **Gains**
  - the knowledge means something to the machine
  - can be mechanically maintained/verified/tuned/learned
  - suitable for general-purpose reasoning methods
There is no middle way

• Every step we take towards objective, public representations takes us farther away from the power and potential of subjective representations

• Public representations are good for everything except AI
Subjective doesn’t mean you can’t build it in

- Subjective ≠ learned
- You can build knowledge in, but you must build it in subjective terms rather than in public, consensual, “objective” terms
- The subjective must be there
Summary

- Subjective experience is the data of AI
  - it’s crazy to try to do AI without experience
- Subjective (predictive) knowledge is powerful
  - automatically verifiable, tunable, extendable, learnable
  - explicit, machine-readable semantics
  - can be built in
- Abstraction is key – in state and time
- Grounding knowledge is a grand challenge