Philosophy 395 Computer Science 228 Symbolic Systems 210	**Lecture 1 — Introduction**	Philosophy of AI Stanford University Winter Quarter, 1989–90

File path: Koyaanisqatsi : Philosophy of AI : Phai — Syllabus

Last edited Monday, January 22, 1990 — 1:30 p.m.

A. Introduction

a. Welcome!

a. Background

a. It’s hardly news that, within a single generation, computers have come to permeate just about every aspect of life — from science, finance, and math, to education, psychology, even the arts.

a. Nor has another prospect escaped the popular: that computers may some day develop genuine intelligence — equaling, or even surpassing, that of humans.

a. The project to develop such machinery, generically called *artificial intelligence*, was officially inaugurated at a famous Dartmouth conference in 1959. In the intervening 30 years, it has given rise to what is almost an independent intellectual discipline: a branch of computer science, some would say, but also connected to psychology, linguistics, education — and philosophy. Hundreds of millions of dollars, departments in the best universities, products (and hype) on Wall Street. Definitely a major undertaking.

a. As John Haugeland points out (at the beginning of his *AI: The Very Idea*), people have constructed machines “in their own image” for centuries (clockwork dolls, e.g.). In this case, however, there’s a difference.

— Historically, the attempts were little more than play — paultry imitations of the human case, which serious people at best took as suggesting, caricaturing, or symbolizing the genuine article.

— In the computational case, however, a much more radical claim is at stake: not that computers might simply *imitate* intelligence, but that computers may actually be able to *have* intelligence.

— Furthermore, the prospect is taken seriously. Some of the best and brightest intellects in the world have dedicated their lives to the pursuit of this dream.

— And I think they’re right. AI, in my view, is by far the most exciting project in late 20th century intellectual life — for deep, interesting reasons.

a. That aim of this course will be to look at the conceptual foundations of this enterprise. Not at technical details, or specific architectural proposals, but at the very foundations on which the project is based. At underlying assumptions, at criteria of success, at plausibility and intellectual merit.

  a. Timely, too; increasing amount of public debate
   — *SciAm* articles: Searle & Churchlands
   — Recent books: Howard Gardner, *Improbably machine*, William Penrose's *The Emprorer's New Mind*, etc.

a. Assumptions

  a. I will assume that people know some AI, and are familiar (in both a practical and theoretical sense) with the computation on which it is based. Courses on compiler design or expert system construction aren't necessary. What you should be familiar with are simple programs, interpreters, and data structures; Turing machines, a touch of computability theory, the notion of a digital state machine. And you should at least have some familiarity with specific architectural projects: theorem-proving, knowledge representation, connectionism, problem solving paradigms, etc.

  a. I will also assume, though less critically, that people are familiar with philosophical investigations and argument. A course in the philosophy of language, mind, or at least a philosophical course on logic would be good. I can imagine someone doing well for whom this is the first course in philosophy, but they will have to work (and think) very hard.

A. The nature of the enterprise

a. In due course, we'll look at various critiques of AI — Dreyfus, Taylor, Winograd, Searle, and others. And, in discussion later in the course, I'll even entertain some of the Big Questions: whether a machine could be conscious; whether a machine could have rights; whether someone who typed ^C at a program could be indicted for murder. But before there's any merit in taking up such questions, we must have the assumptions and overall structure of the field in much clearer view. So that's will be the goal of this first lecture: to set out the lay of the land in which AI is conducted.

a. Strong vs. weak AI.

  a. Start with a basic dichotomy, using terminology introduced by John Searle. To get the discussion going, I'll assume (this is rough) that the goal of AI is to use computers or computational vocabulary to develop or understand intelligence. These are rough-and-ready characterisations; crisper ones will emerge in a moment.

  a. Right away, two possibilities can be discriminated:
   i. **Weak AI**: using computers to model intelligence.
   ii. **Strong AI**: developing computers that actually *have* intelligence.

  a. A couple of words on each.

———————————————————————————

a. Weak AI
  - — like thunder storms, tectonic plate movement, etc.
  - — isn't a claim that the *subject matter* is computational at all
  - — so what's the computer for?
    - — various possibilities: pen & pencil, telescope, calculator
    - — quotes
      - — L&F
      - — Johnson-Laird (new book)
      - — Edelman (NYTimes)
  - — obviously gives much more freedom
  - — but right away there are a whole spate of problems:
    - i. if AI is distinguished at all (isn't just philosophy of mind, or synthetic biology), what it is to be a computer still needs to be addressed
    - ii. role of the computer
    - iii. if it is *modelling*, then what it is to be a model (get back to that).
    - iv. "implement the theory". Spectacular claim (depends on what implement means).
    - iv. what is subject matter: intelligence, perception, human affairs independent of what they are, etc.
    - v. Now real confusion: morale:
      - • Theory, model, and subject matter are all of approximately the same *type*. (contrast Kepler's stars, brass orrery, and hand-written claims).
      - • will have important consequences
    - v. Summary
      - — intellectual IOU's: (model, theory, implement, intelligence, computational, …)
      - — confusion of types
      - — threat of vacuity

b. Strong AI
  - — strength
    - — At least it has some bite
    - — Radical, too. Means that people — at least in being intelligent, or thinking, or some such (still haven't said exactly what) — aren't at the center of the universe. What Kepler did physically, AI may do intellectually.
    - — People are computers. Don't beat around the bush.

  - Furthermore, this is real people: friends and lovers. So: if you write something, and fancy yourself in strong AI, be prepared to treat your dearest acquaintances as if what you write is true of them.
- But look at this is some detail. After all, some of the IOU's from weak side are still lurking in the shadows.
- Rough equivalence (imply a larger, encompassing class)

    ⇒ **The computational claim on mind**

  a. Computational side
  b. Human side

- Computation
  - various theories
    - **fsm**: formal symbol manipulators
    - **dsm**: digital state machines
    - **rft**: effective recursive function calculators (algorithms)
    - **ip**: information processors
  - all conceptually different
  - will look at them all
  - (I believe they're all wrong; but not for now)
  - empirical question!
    - therefore judge: practice
  - problem: most people who engage in the debate *aren't computer scientists*
    - quote Fodor: to run, must be compiled.
    - entirely vulnerable to what they take computers to be.
- Human side
  - must be in virtue of some human property
  - not mass, or sexual reproduction, or evolutionary history
  - what makes the equation plausible?
  - thinking, calculating, beliefs, etc. mind. symbol
  - i.e., what recommends: symbols, language, etc. ⇒ intentionality
  - leads to: what I will call *intentionality*.
    - word is from Brentano. Cf. Searle's book.
  - slide of what it includes
- Further complication
  - Not *whole* to *whole*. Must be a part.
  - So what part.
  - ⇒ particular architectural proposals
    - logico-deductive
    - problem-solving/search

    - connectionist
    - procedural
- Summary
  - computational side: various possible theories (enumerate)
  - human side: intentionality (language, problem solving, mind, etc.)
    - real people we're talking about
  - sub-types in each case
    - what is the type, and what the sub-type
  - ϕs worry about *whole* of comp; AIer's, about *species*

Theoretical Structure

- Talked about computers offering promise, but really two
  - *practical* promise: manifesting intelligence
  - *intellectual* promise: of providing the wherewithal to *understand* intelligence (— independent distinction from weak/strong)
- In this course, I will be more interested in the second
- But spell it out a little, in terms of this identification of intentionality
  - Offer the promise of explaining, in *non*-intentional terms, one of the outstanding questions of intellectual history.
  - Note: this is a generally interesting problem
  - Other routes in:
    - biology (Milikan, others)
    - physics (entropy, Bohm, etc.)
    - consciousness (Searle)
- So where do we stand?
  1. Computation: empirical question
     - Note: might not be a subject matter (cf. cars)
  2. Strong AI: people, in virtue of their intentionality
     - That too is an empirical question
     - People may not be a subject matter either, of course (probably aren't)
  3. What's at stake; what can be assumed
     - Cf. slide

Prospectus

- In subsequent lectures, …
  - review all 7 lectures
  - as much ϕ of computation as ϕ of cognitive science — defend
  - goal: to Educate (not teach my philosophy of such things). Up the ante on public discussion. Know the issues; be able to assess others' contributions (e.g. in recent SciAm articles).

- Other
    - prerequisites
    - structure: lecture/discussion, exams, etc.
    - dates of make-up classes
    - enormous amount of material; only cover a bit

- **Slides to prepare**

    a. Intentional terms
    b. Things that may *not* be assumable (different depending on different writers)

- **Points to be made**

    0. Most exciting intellectual project this century, perhaps next as well.
    1. This is people we're talking about: hold true of friends & lovers. Criterion of *humanity* .
    2. No distinction in type between theory, model, and subject matter.
    3. Strong vs. weak AI. Intentional assumptions in weak. Not as clear in strong! Computational claim on mind.
    4. Plausibility of strong AI entirely dependent on underlying model of computation. *Empirical* criterion.
    5. Under strong: different models of computation (FSM, RF, DSM, IP, etc.).
    6. What project? What's to be explained? What's to be assumed? Methodological assumptions. Naturalism; conceptualism. Betrayal of one's metaphysics (cf. B&P).
    7. Assumptions about people: theories of psychology — goals, behaviourist, mental, BDI, etc. (especially relevant under the construal that it is to provide a *foundation* for psychology).

- **Morals**

    - Computation: we don't understand. Empirical inquiry. Do justice to practice.

- **Notes**

- **Quotes**

    - "According to strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations." [Searle's MB&P, 1st ¶]

——end of file ——��