Guide 14

1. AI: Philosophy and ethics
   1. What are the basic philosophical questions regarding AI?
      1. Is artificial general intelligence possible?
         1. Can a machine solve any problem that a human being can solve using intelligence?
         2. Are there hard limits to what a machine can accomplish?
      2. Are intelligent machines dangerous?
         1. How can we ensure that machines behave ethically and that they are used ethically?
      3. Can a machine have a mind, consciousness, and mental states in exactly the same sense that human beings do?
         1. Can a machine be sentient, and deserve certain rights?
         2. Can a machine intentionally cause harm?
2. Alan Turing’s “Computing Machinery and Intelligence” — Read Sections 1–3; 6—7, skimming the “contrary” arguments to get their basic thesis.
   1. What actual, operational question does Turing address in the paper?
      1. What will happen when a machine takes the part of the A – a man, in this game?
   2. Specify the *imitation game*.
      1. Three players: a man, a woman, and an interrogator (either sex).
      2. Interrogator is isolated physically from the man and the woman.
      3. Goal: Interrogator needs to determine who is the man and who is the woman.
      4. Interrogator is allowed to question the man and the woman.
      5. The man is supposed to try to mislead the interrogator.
      6. The woman is supposed to try to lead the interrogator to the correct conclusion.
   3. Do any of Turing’s “contrary” views seem compelling to you?
      1. Theological objection – machine doesn’t have a soul.
      2. “head in the sand” objection – too terrible to think machines are sentient.
      3. Mathematical objection – Godel’s Theorem – limits to the power of any system but no limit to human intellect.
      4. Argument from consciousness – machines cannot feel or express emotion or produce art.
      5. Argument from various disabilities – machine can do all these things but can’t do X.
      6. Lady Lovelace’s objection – machines can never do anything new, only what humans have already done.
      7. Argument from continuity in the nervous system – finite (discrete) state machine cannot mimic the human nervous system (continuous state machine).
      8. Argument from informality of behavior – impossible to model every single possibility, hence machines can never be like humans.
   4. Does Turing give any prescriptions for how one might program a computer to learn?
      1. Simulate the mind of a child rather than an adult mind.
         1. Subject to course of education to obtain the adult mind.
      2. The mind of a child is full of blank sheets with little mechanism that makes it easily programmed.
      3. Divide into 2 parts:
         1. Child programming.
         2. Education process.
      4. Use the principle of punishments and rewards in the education process.
      5. Implement a system of logical inference.
      6. Include a random element in the learning machine – useful for problem solving.
      7. Refer to Section 7 of his essay for further details.
      8. For further reading, Sections 4–5 give a review of computers and Turing machines.
3. John Searl’s Chinese room (Introduction and Sections 1–3)
   1. What is Searle’s basic argument as it relates to the Turing test?
      1. Question to ask:
         1. Does the machine literally understand Chinese?
         2. Or is it merely simulating the ability to understand Chinese?
      2. Given a computer that can pass the Turing Test by convincing a human Chinese speaker that it itself is also a human Chinese speaker…
      3. Argument:
         1. There is no difference between a human running the program “manually” and the computer running the program to produce Chinese characters as output in response to the Chinese characters input given by the human Chinese native speaker.
         2. The human and the computer both follow a program step-by-step to produce behavior that is interpreted by the user as demonstrating intelligent conversation.
         3. However, the human doesn’t actually understand Chinese and hence neither does the computer.
         4. Without understanding (intentionality), we cannot describe what the machine is doing as thinking and since it does not think it does not have a mind in anything like the normal sense of the world.
            1. Therefore, strong AI is false.
   2. Compare and contrast *strong* vs. *weak* AI
      1. Strong AI: literal understanding.
         1. “the correct simulation really is a mind”.
      2. Weak AI: simulated understanding.
         1. “the correct simulation is a model of the mind”.