

From Computation to Thinking — Teaching Guide

A guide for teachers and reading groups: six sessions, one per part of the book, each with its topics, reading time, discussion questions and a short activity.

The book in brief

From Computation to Thinking explains how a machine that only moves symbols comes to answer a frightened parent's message at 2 a.m. One example runs through the whole book: a parent writes about a feverish child, and an AI nurse must decide what to do. Each chapter climbs one rung, from rule-following machines to learning, prediction, reasoning and agents. The last parts ask what the machine is, and how it fits into a hospital team.

- **Who it is for:** curious readers with no technical background; students in introductory AI, AI-in-healthcare and nursing-informatics courses; professionals who use these systems at work.
- **Prerequisites:** none. The book has no mathematics.
- **Reading time:** about three hours; most chapters take 6 to 13 minutes.
- **Built-in material:** every chapter opens with its main idea in one sentence. The glossary defines the technical terms, and the notes at the back list the studies each chapter draws on.

Session plan

Session 1. Machines that follow rules (Part I, Chapters 1-3)

- **Topics:** symbols, states and rulebooks; Turing machines; universality; the limits of computation (undecidable questions, problems too slow to solve).
- **Reading:** about 20 minutes.
- **Discussion:**
 - If meaning is supplied by the reader of the tape, where would a machine's own meaning come from?
 - A language model's weights stay fixed while its prompt changes. Is the prompt a program?
 - Checking is easy and finding is hard. Is "judging is easier than doing" the same idea?
- **Activity:** run the book's add-one machine by hand on a strip of paper. Then ask: which step "knew" it was adding?

Session 2. Machines that find rules (Part II, Chapters 4-7)

- **Topics:** expert systems (MYCIN); learning from a measure of mistakes; generalization and why it can work; features learned for seeing and hearing; shortcuts that are right for the wrong reason.
- **Reading:** about 30 minutes.
- **Discussion:**
 - If an expert cannot say what she knows, in what sense does she know it?
 - The loss decides what the machine learns. Who decides the loss?
 - Which "labels" in nursing quietly stand in for something else: patient satisfaction for good care, a normal temperature for a well child?
- **Activity:** in small groups, write rules for answering the 2 a.m. message. Swap rules and find a case each set gets wrong.

Session 3. Machines that predict language (Part III, Chapter 8)

- **Topics:** next-word prediction; why predicting well requires knowledge of the world; the base model as every writer at once.
- **Reading:** about 10 minutes.
- **Discussion:**
 - If prediction forces a model to compress the reasoning behind the text, is a good enough predictor already a reasoner?
 - Text records what people wrote, not what was true. How much of the world can be learned from what people say about it?
- **Activity:** read a paragraph aloud and stop mid-sentence. The group predicts the next word. What did they need to know to get it right?

Session 4. From predictor to agent (Part IV, Chapters 9-13)

- **Topics:** imitation (teaching a role); learning from consequences and human judges; written reasoning; agents that act with tools; the machine's scale in books, kettles and cities.
- **Reading:** about 45 minutes; split across two meetings if needed.
- **Discussion:**
 - Imitation teaches the model to *sound* careful. What would it take for it to *be* careful?
 - The judge is a learned imitation of people. Who checks the judge?
 - Who is responsible when an AI nurse with a delegated goal gets it wrong: the model, the harness, the builder or the user?
 - A child learns language from a tiny fraction of the text the model reads. What does the child have that the model lacks?
- **Activity:** give the group two answers to the 2 a.m. message. Ask them to say which is better, then to write a better one. Compare how long each task took.

Session 5. What it is, really (Part V, Chapters 14-16 and the Coda)

- **Topics:** meaning and grounding; hallucination as a failure of correctness; goals, agency and curiosity; uncertainty, calibration and knowing when to stop; whether anyone is home.
- **Reading:** about 40 minutes.
- **Discussion:**
 - People who accept a model's answer without checking it hold a map they have not tested themselves. Is *their* knowledge borrowed too?
 - Which of a nurse's questions could be found by "maximize information", and which need her to care about this child?
 - If a machine's sense of uncertainty can be turned up and down by hand, who should hold the dial: the builder, the hospital or the user?
 - What evidence, if any, could tell a system that has experiences from one that only reproduces the signs of having them?
- **Activity:** ask a chatbot three questions, one of them obscure. Check every claim against a source. Did its confidence match its accuracy?

Session 6. Living with it (Part VI, Chapters 17-18)

- **Topics:** the machine in a hospital team; shared records and consistent documents; alert fatigue; human-AI teams that do worse; which tasks to hand over, decided by whom.
- **Reading:** about 20 minutes.
- **Discussion:**
 - How many alerts can a nurse in charge sort in a night before the sorting itself fails?
 - If drafts reduce burnout without saving time, what exactly did they save?
 - Where will tomorrow's experienced nurses learn their judgment if the machine handles the routine cases they used to learn on?
- **Activity:** list the tasks in one job the group knows well. Mark each: hand to the machine, share, or keep. Who should decide?

Shorter formats

- **A single session:** Chapter 8 (prediction), Chapter 14 (meaning) and Chapter 17 (the team). About 30 minutes of reading.
- **A healthcare group, three sessions:** Chapters 4, 7 and 12; Chapters 16 and the Coda; Chapters 17 and 18.
- **A technical course:** pair each chapter with the studies listed in its notes.

All chapter questions

The session plan selects from these. Teachers can draw on the rest. They come from the chapters' closing questions; three are reworded slightly to stand alone.

1. Moving Symbols

- If meaning is supplied by the reader of the tape, where would a machine's own meaning come from?
- The rules are local, yet the result (adding one to a whole number) is global. How does local action produce global results? The same question returns, much larger, when billions of tiny local updates add up to a fluent answer to a frightened parent.
- The space of snapshots is huge, but the path is thin. What decides which thin paths are *useful*? That is the problem learning will try to solve.

2. One Machine for Every Rulebook

- If universality is so cheap that very simple systems have it, what makes one universal system more *useful* than another?
- A language model's weights stay fixed while its prompt changes. Is the prompt a program?
- A brain cannot copy its knowledge the way a computer copies a file. Is that a weakness, or the price of its efficiency?

3. The Walls

- Checking is easy and finding is hard. Is "judging is easier than doing" (Chapter 10) the same idea?
- If intelligence lives in "pockets of reducibility", how does a system find them?
- Contrary defeats any checker by asking about itself. Does anything like that happen when an AI system is asked about its own behavior?

4. Writing the Nurse's Rules by Hand

- If an expert cannot say what she knows, in what sense does she *know* it? Could a machine know something in that way?
- MYCIN could explain every conclusion. The learned systems that replaced it mostly cannot. What was lost?
- Which parts of nursing *can* be written as exact rules, and should stay that way?

5. Learning: Finding the Rules Instead of Writing Them

- The loss decides what the machine learns. Who decides the loss?
- If a network's rule is a billion numbers, in what sense does it "have" a rule at all?
- Training is a path through a space of weight settings. Do different starting points end in different rules, and does it matter?

6. Why Learning Can Work at All

- If every learner needs assumptions, what are a human child's built-in assumptions, and where did they come from?
- Is the second student's principle "true", or merely shorter? Could a shorter description be wrong?
- Scaling laws are smooth. Abilities sometimes seem to appear suddenly. How can both be true?

7. Learning to See and Hear

- If a network can't tell us which features it uses, how would we ever find the purple pen before it matters?
- A good nurse also relies on cues she can't fully state (Chapter 4). How is her tacit knowledge different from a network's shortcut?
- Which "labels" in nursing quietly stand in for something else: patient satisfaction for good care, a normal temperature for a well child?

8. Predicting the Next Word

- If prediction forces a model to compress the reasoning behind the text, is a good enough predictor already a reasoner?
- The base model can continue any document in any writer's voice. Is there anything in it that is "the model itself"?
- Text records what people wrote, not what was true. How much of the world can be learned from what people say about it?

9. Imitation: Teaching It a Role

- If the base model already "knows" the emergency, where exactly was the knowledge, and why could it not use it unprompted?
- Imitation teaches the model to *sound* careful. What would it take for it to *be* careful?
- The model plays a role. Is there anything underneath the role, or is the role all there is? (Chapter 14 returns to this.)

10. Learning from Consequences

- If judging is easier than doing, can a machine that judges well eventually do better than its judges? Where would that stop?
- The judge is a learned imitation of people. Who checks the judge?
- Which parts of a good nurse's answer could ever be scored exactly, and which only by comparison?

11. Thinking Out Loud

- If the reasoning habits were present "sometimes" before training, where did they come from? From the people whose writing the model read?
- What would a reliable checker for a nurse's explanation look like? Could it exist?
- Copying is cheap and discovery expensive. Who will do the discovering, and on which problems?

12. Agents: Closing the Loop with the World

- If framing a task differently changes what a model learns from it, is something in the model playing the part of a self-description? How could we tell?
- Who is responsible when an AI nurse with a delegated goal gets it wrong: the model, the harness, the builder, or the user?
- How would you test an AI nurse in "settings that look like real use" without putting a real child at risk?

13. How Big Is It? The Machine in Numbers You Can Picture

- A child learns language from a tiny fraction of the text the model reads. What does the child have that the model lacks?
- If answering costs about as much energy as your brain spends reading the answer, where is the real cost, and who pays it?
- Most of the energy goes into moving numbers around, not into arithmetic. What would a machine look like that moved far less data?

14. Meaning: Are Its Words about Anything?

- Harnad says only *some* symbols must be tied directly to the world, and the rest can be built from them. How many anchors would a language model need?
- If a model could check every claim it makes through tools, would its meaning still be "borrowed"?
- People who accept a model's answer without checking it hold a map they have not tested themselves. Is *their* knowledge borrowed too?

15. Agency and Curiosity: Whose Goals, Whose Questions?

- If a goal written in text is often easier to game than a goal felt by a body, is there anything in between? And how could the comparison be tested?
- Which of a nurse's questions could be found by "maximize information", and which need her to care about this child?
- Should we want machines that choose their own questions? Who would choose what they care about?

16. The Walls Again, and Knowing When to Stop

- If a machine's sense of uncertainty can be turned up and down by hand, who should hold the dial: the builder, the hospital, or the user?
- Which parts of a good nurse's judgment can ever be checked cheaply, and which never can?

Coda. Is Anyone Home?

- If a machine came to look after its own memory, tools and resources, would its meaning stop being borrowed? Would that change anything about whether anyone is home?
- What evidence, if any, could tell a system that has experiences from one that only reproduces the signs of having them?

17. The Machine on the Ward: Working in a Team

- If a machine can draft every document in a case, who should be allowed to correct the record, and how would the correction reach every copy?
- How many alerts can a nurse in charge sort in a night before the sorting itself fails?
- If drafts reduce burnout without saving time, what exactly did they save?

18. Replace or Relieve? The Nurse and the Machine

- If a hospital can choose between the two futures, what rules, incentives or votes would push it toward the second?
- Where will tomorrow's experienced nurses learn their judgment if the machine handles the routine cases they used to learn on?
- If patients prefer the machine's written answers, is that because the machine is better, or because the humans were never given time to write well?