Qualia of the World and the Nature of Consciousness



Eudgene Brenev
Mira Veyran

2026

Qualia of the World and the Nature of Consciousness

A Practical Textbook
From Observation to Models and Design

Eudgene Brenev

Mira Veyran

From Personal Experience
to Scientific Inquiry and Design

2026

Preface

Pick up a familiar cup. You feel its weight, notice its warmth, and see the color of the glaze. Perhaps the smell of the drink brings back a memory. All of this happens within a few seconds. Yet if you pause over this simple episode, questions begin to arise. What makes up an experience? Why does it change? How do we know what we feel and think? What would have to happen in a living or artificial system for it to have experience of its own? Our book begins with these questions.

The word "qualia" refers to a familiar aspect of life: what it is like to experience something. Seeing red, hearing a quiet sound, or feeling a rough surface means encountering qualities of experience. We will learn to notice differences between them and find precise words for them. Questions about consciousness, the sense of self, and reflection will gradually enter the discussion. We will examine the relationships among experiencing an event, being able to report it, controlling one's actions, and thinking about how one thinks.

You can begin without a background in philosophy, neuroscience, or programming. The first observations require only a notebook. The research sections will gradually call for probability, causal comparisons, and reading code. Definitions and worked examples precede these tasks. Developing a hypothesis of your own comes after you have mastered an existing example; it maintains the book's demanding scientific standards.

We will begin with careful description. This takes practice: a single sentence can easily mix a sensation, a judgment, and an assumption about its cause. We will then consider self-awareness: what it means to experience a body as one's own, attribute an action to oneself, or notice one's own thought. Reflection adds the ability to return to a decision and ask what it was based on. Keeping your early notes will help you see how your explanation changed and what information became available only after the event.

From description, we will move to inquiry. What should we compare or measure, and which conditions should we change? How can we tell which of two explanations fits a result better? We will design small tests, work with uncertainty, and record our expectations in advance. A useful question for any idea is: what result would make me reconsider it? Finding a weakness in your explanation can help you frame the next experiment more precisely. Observation thus gradually becomes the basis of a testable hypothesis.

Exercises and code will help reveal how a proposed mechanism works: how traces of the past are retained, signals accumulate, and predictions change. You can run a small program step by step and understand why it produced a particular answer.

Later, we will encounter systems that perceive their surroundings, choose actions, use memory, and evaluate decisions. Changing one element will let you observe what depends on it. Existing examples are enough to get started. Later, code will become a way to express your own ideas so that someone else can reproduce the calculation.

Along the way, we will encounter scientific theories and research on consciousness, self-control, and perception. We will look at how animals are studied, artificial systems are discussed, and biological and hybrid models are designed. In each field, we will clarify what has been observed and what connection to experience its authors propose. We will seek conditions under which competing ideas make different predictions. This makes progress possible even where no final answer exists: we can formulate a clear question and a way to test it.

The aim of this book is to help you develop your own hypotheses about the origin and organization of experience, then move toward designing systems that might be considered candidates for subjecthood. Here, a subject means a presumed bearer of experience. The book does not offer a proven way to create one. We can clarify what properties such a system should have, why we associate them with experience, and what would support or weaken that association. A working model helps us investigate these questions and identify what grounds are still missing.

The best way to read this book is with a notebook. Record an expectation, perform the exercise, compare the outcome with your prediction, and answer the questions before reading the answers. Keep both successful solutions and difficulties: a research project of your own can grow out of them. By the end of this path, you will be better able to describe experience, analyze explanations, formulate testable questions, assemble simple models, and justify the next step. We hope this book will help you begin that work and stay curious about what remains unexplained.

Eudgene Brenev and Mira Veyran

How to Work with This Book

This book begins with a question anyone can ask: what exactly am I experiencing right now? By its end, that question will have developed into a research question: what mechanisms might explain the organization of experience, and how can we design a system for which the assumption of subjective experience has some definite basis? Between these points lie observations, concepts, measurements, scientific theories, and working models.

We will use "qualia" for the qualitative aspect of experience: how a color appears, a melody sounds, or an action feels. This working usage does not commit us in advance to a single theory of the nature of experience. When the author of a scientific paper assigns additional properties to the term, we will examine them separately. In the engineering sections, a subject is a presumed bearer of experience. Self-awareness and reflection raise questions of their own and require their own methods of inquiry.

For the opening chapters, you need only curiosity, an ordinary notebook, and a willingness to refine your descriptions. Mathematics and programming appear gradually, after the problem has been explained. The practical sections include ready-to-run programs: you can first observe a mechanism at work, then change its parameters, and finally read and extend the code. The biological track also addresses trained specialists and includes data analysis, computational models, and research design.

The Path through the Book

Chapters 1–4 teach you to notice experience, distinguish its properties, and track your own understanding. Chapters 5–8 introduce explanation, measurement, causal testing, and working with uncertainty. Chapters 9–12 bring in neural networks, feedback, the body, memory, and the evaluation of thought. These sections establish a language for discussing scientific theories without merely memorizing their names.

Chapters 13–16 compare theories and contemporary ways of testing them. Chapters 17–20 distinguish several meanings of emergence: the onset of an episode, individual development, evolutionary origins, and the basis of experience's very existence. Chapters 21–23 consider animals, systems without neurons, and artificial candidates. Keeping these questions separate helps prevent an answer to one from being presented as a solution to all the others.

In Chapters 24–27, we develop an explanation of our own and give it substantive rivals. Chapters 28–32 move to software, biological, and hybrid implementations.

The final four chapters help turn a model into a research project: define a formalization, describe the test in advance, analyze the result, and prepare a coherent account.

Work through each chapter with written notes if possible. First state what you expect from the exercise. Then perform it and separate the outcome from your explanation. At the end, answer the questions before reading the answers. If your conclusion differs from the worked discussion, check whether the initial conditions changed. Sometimes a discrepancy reveals a promising new problem.

What to Keep in Your Research Notebook

For an observation, record the situation, time, and details you noticed. For an experiment, add the conditions, outcome measure, and initial expectation. For a program, save the command, parameters, and result. When analyzing a paper, distinguish the data from its authors' conclusions. Keep the original record: it helps separate a new observation from a later interpretation. A model's version must also remain linked to its result.

An exercise is complete when you can explain exactly what you did and what the result permits you to conclude. Confidence and a detailed account do not replace testing. At the same time, uncertainty does not mean that the work is useless. A good result may show that two explanations cannot yet be distinguished and identify the conditions for the next experiment.

Reading Claims of Different Status

The text includes working definitions, findings from published studies, theoretical propositions, and the authors' hypotheses. References in square brackets point to the bibliography. A particular experiment concerns a particular sample and task. A theoretical conclusion depends on explicitly stated assumptions. The authors' models are proposals for testing; describing them does not make them discoveries.

Numerical teaching examples are explicitly identified. The practical results were obtained by running the supplied code and describe constructed environments. Their relationship to experience requires a separate argument.

The illustrations follow the style of a scientific atlas. First read the figure title, then trace the indicated relationships or compare the panels, and only then consider the conclusion in the caption. Graphs identify the quantity, conditions, and provenance of the data. Teaching diagrams, program outputs, and research hypotheses rest on different foundations. Color helps you recognize elements; labels determine what they mean. The index at the end of the book lets you return to a particular figure.

Why the Book Has an Engineering Goal

Trying to build a system forces an explanation to become definite. You must decide what is stored, what changes, what information is available, and what action follows. A vague idea can often accommodate any story after the event; an executable model makes specific commitments. Even its failure can clarify what mechanism is missing.

The engineering task also includes a deeper question: why should the chosen organization have anything to do with experience? We will not evade it by substituting an impressive program report for experience. The research sections examine possible principles linking the two, the arguments for them, their invariants, and the difficulties of testing them. Success requires connecting a precise mechanism with a serious explanatory question.

Scientific work allows us to make assumptions of our own. It requires us to make them clear enough for someone else to check the calculation, challenge its basis, and propose a discriminating experiment. That is why the book ends with a research project. It gives the reader a way to continue working where no ready-made answer exists.

Keeping the Thread of a Chapter

Each chapter opens by recalling an earlier result and explaining the new question. The main text introduces concepts and scientific arguments. Before the practical exercise, there is an additional worked case, followed by a brief pause to establish what is now known. The ending leads to the next question. If you lose the thread, return to that pause and state in your own words the result you will need next.

The practical exercises have a distinct format. First read the objective and prerequisites, then follow the numbered steps and save the specified result. A hint helps you recover the next step. The research extension calls for decisions of your own and may take substantially longer. Time estimates refer to the basic exercise and are not a standard for how quickly you should work.

Answers to the self-check questions and discussions of all the exercises are collected in Appendix F. Open them after making an attempt of your own. For observations and research projects, the example explains quality criteria; it does not prescribe the only permissible experience or theory. A negative result or a tie between rivals can also be a substantive outcome.

Three Routes through the Same Material

A first sequential reading: Read every chapter in order and complete the basic level of the exercises. For each new concept, keep its definition and an example of

your own. After each part, take a break and reread its transition. Scientific arguments and qualifications are part of this route.

Reproducing mechanisms: After Chapters 5–12, use the existing calculations in Chapters 13, 15, 22, and 25, followed by the code in Chapters 28–30. Reproduce the baseline run before changing parameters. Chapters 33–35 cover reconstructing mechanisms, the protocol, and deciding what follows from the result.

A research project of your own: Work through the basic comparisons, then return to Chapters 24–27. State the mechanism, its connection to experience, its closest rival, and a possible falsifying result separately. Chapter 31 extends the work to real biological data. Chapter 36 helps you write up work that can be tested.

Prerequisites for the More Advanced Sections

Section	Preparation	What You Gain
11 and 15	Probability and likelihood; normalization in Chapter 11	Predicting the world and the effect of one's own actions
13 and 14	Intervention and measurement from Chapters 6–8	The distinct commitments of different theories
22 and 23	A strong rival; transfer and controls	A candidate's profile before programming begins
25 and 26	An update rule; sufficiency and necessity	Executable memory and a separate principle concerning experience
28–30	Specifying inputs, states, and permitted information	Reproducible mechanisms and ablation studies
31	Grouping and a baseline predictive model	A testable analysis of an open data table
33–36	Distinguishing models and paired comparisons	A protocol, a decision, and a research account

Contents

Preface 3
How to work with this book 5

Part 1 From experience to self-awareness and reflection 11
1 What exactly are we experiencing? 11
2 Differences and similarities of experiences 18
3 What It Means to Experience Oneself 26
4 How we notice our own thinking 33

Part 2 Transition to scientific research 39
5 What counts as an explanation? 39
6 How to Explore Subjective Experience 45
7 How the check works 52
8 How to deal with uncertainty 58

Part 3 Mechanisms and models 65
9 From nerve cell to system 65
10 Signal processing and access to their content 71
11 Body and action in self-reference 77
12 Memory and time in assessing thinking 85

Part 4 Theories and cutting edge research 92
13 Workspace and recurrent processing 92
14 Integrated information and higher order theories 99
15 Prediction and self-models 106
16 Where data differentiates between theories 112

Part 5 The emergence of consciousness and self-awareness 118
17 How does a single episode occur? 118
18 How self-awareness develops 124
19 How to Explore the Evolution of Experience 131
20 What an Explanation of Emergence Requires 138

Part 6 Possible bearers of experience 144
21 Consciousness and self-control in animals 144
22 Boundaries between adaptation and experience 152
23 Artificial and biological candidates 158

Part 7 Developing your own explanations 165
24 Requirements for the future model 165
25 From idea to mechanism 171
26 From mechanism to experience 178
27 Competitors and risky predictions 186

Part 8 From words to practice 193
28 An Agent in an Environment with a Verifiable Task 193
29 Perception and Global Access 202

30 Self-model and reflection based on memory 210
31 Biological and hybrid implementations 220
32 Verification of the designed system 229

Part 9 Research work 237
33 Formalization and simulation 237
34 Protocol before data collection 245
35 Review and revision 252
36 Your Own Research 259
Appendix A Working Dictionary 266
Appendix B Launch and arrangement of the workshop 271
Appendix C Paired comparison of results 274
Appendix D Forms for independent work 275
Appendix E Index of illustrations 277
Appendix F Reviews of practices and self-tests 279
Sources 298
Author and projects 301

Part 1

From experience to self-awareness and reflection

Chapter 1

What exactly are we experiencing?

Let's start with a small skill that will be needed throughout the book: distinguishing between the event itself, the description of the experience, and the explanation of its cause. To do this, you don’t need to look for unusual conditions. One touch is enough. We are not yet building a theory of consciousness; we are preparing material about which a future theory will have to explain something. By the end of the chapter, you will have your first record and understand which parts of it require different methods of verification.

Start of research

There is a mug on the table. You can name its material, approximate volume, and price. Now take it in your hand. Other descriptions become possible: the handle feels cool, the rim seems smooth, and a bright reflection is visible on the surface. The object remains part of the same situation, but the question has changed. We are interested in what encountering it feels like.

Thus begins the study of qualia. The first work will not require an unusual condition or a rare experience. Something that is already happening when you look at a page, hear a sound, or feel a touch will do. Our task is to turn a familiar, usually quickly passing event into material about which we can talk accurately. By the end of the chapter, you will have your first observation record and several questions to test later.

What does the word qualia mean?

In this book, qualia refers to the qualitative characteristics of subjective experience: the way something is felt or perceived. Coolness is different from roughness, and a sharp sound is different from a dull sound. The difference exists for the experiencing person, even if he has not yet found the appropriate words. We choose a broad working meaning of the term in order to first learn to identify the subject of research.

The word "subjective" here indicates the presence of the experiencer's perspective. It does not mean "arbitrary" or "necessarily erroneous." You can report fairly confidently that the touch felt cool and be mistaken about the temperature of the object. These statements refer to different properties of the situation. The temperature can be compared using the readings of the device; the feeling of coolness describes what happened in your experience.

Qualia are not limited to emotions. Emotional tone can be part of an experience, as can color, sound, or bodily sensation. At the same time, there is no need to turn every thought into a set of imaginary pictures in advance. If, while solving a problem, you notice words, a vague feeling of difficulty, or an expectation of an answer, describe exactly this. The content of the observation is more important than conformity to someone else's model.

In philosophy, stronger meanings of the term are also used. Daniel Dennett has criticized the idea of qualia as special properties that are attributed, in particular, to ineffability and unmistakable accessibility to the person himself. His argument is directed at a particular conception of these properties. In our work, the presence of a verbal description will neither be a refutation of the experience nor a guarantee of complete accuracy of its understanding. [1]

Event, experience and recording

Consider the phrase: “I touched the handle, felt the coolness and realized that the mug was standing near the open window.” In one sentence there was a report of an action, a description of an experience, and an assumption of a antecedent cause. If left merged, it is easy to mistake the entire sentence for direct observation.

Separation helps you see what exactly needs checking. The touch can be recorded on video. Coolness is available in the person's report. An open window is a separate version of the origin of temperature, for which other information may be needed. Even a plausible explanation remains an explanation; it does not become part of what is seen just because it arose quickly.



Figure 1. Three areas of description of one episode. The event, the description of the experience, and the hypothesis about the cause refer to the same situation but answer different questions. On the right are teaching statements; the image itself does not establish the feeling of coolness or its physical cause.

Statement	**What does it say?**
I turned the mug	Action in a situation
The highlight has shifted to the right and become brighter	Description of visual change
The mug is well made	Judgment about an object
The coolness weakened because the handle got hot	Description of change and causal hypothesis

The words “pleasant” and “unpleasant” are not automatically distributed across this table. “The touch felt good” can inform the emotional tone of the episode. “This is a nice mug” may express a habitual preference or a general impression. You need to clarify the meaning: what happened now, what you usually prefer and what conclusion you draw about the thing.

Why is a detailed report not everything?

Suppose you first wrote down: “The mug is cool.” Then it turned out: “At the point of contact there was a feeling of coolness, which became less noticeable towards the end.” The second recording more accurately localizes the sensation and highlights its change. This is an improvement in the way of description. However, the number of words does not indicate how rich the original experience was.

There are several possible paths to a longer entry. You could take a closer look at the subject, find the right word, remember a detail, or include an assumption in the text that was not there before. Therefore, it is useful to save an early recording. It allows you to compare not only the conclusions, but also how they came together. Clearly corrected text often hides the origins of the confidence.

Time also matters. Recording during an event, immediately after it, and in the evening from memory are different research situations. In all three cases, the person can provide valuable information. But in the latter case, he describes the past through a present memory. A simple note “recorded in about a minute” makes the material clearer to the future reader and yourself.

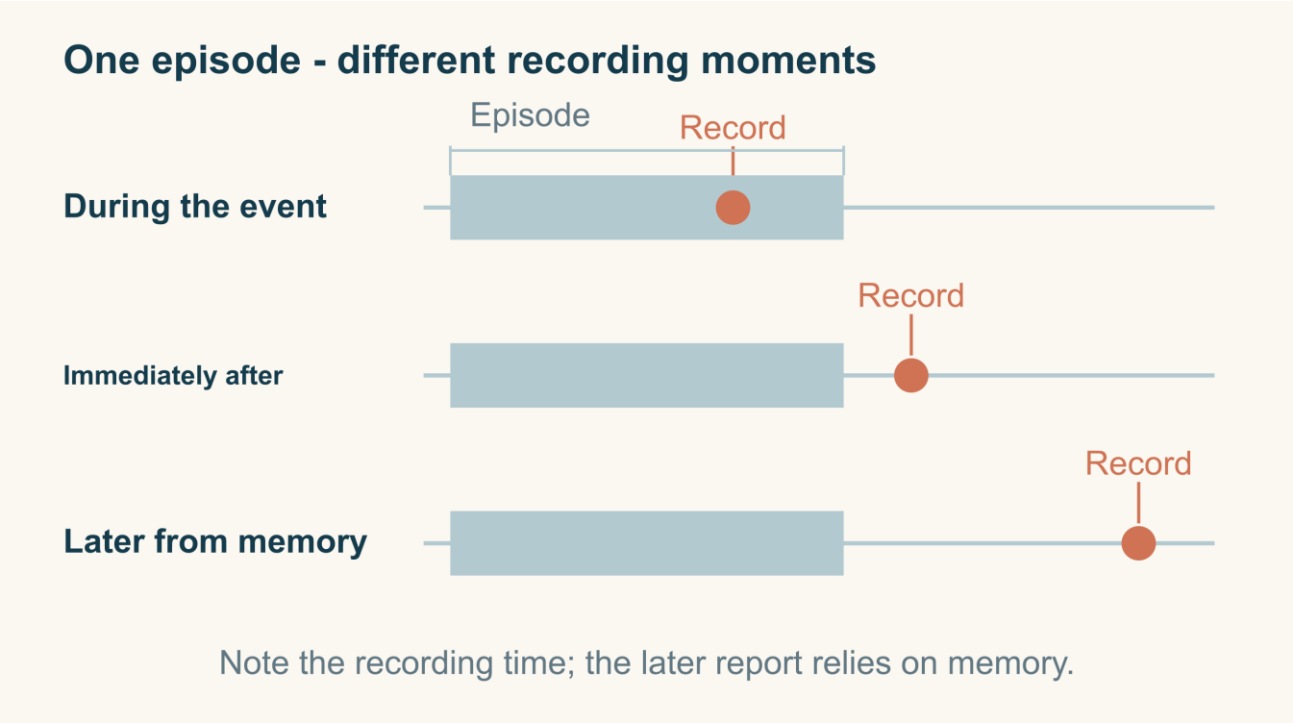

Figure 2. Recording time as part of observation. Descriptions during the event, immediately after it, and later from memory are collected in different conditions. Indicating the moment of recording allows you to distinguish the current report from the memory and subsequent explanation.

Where Do Consciousness and Reflection Fit?

The word "consciousness" refers to several related issues. Does the system have any subjective experience at all? What information is available to it to respond and act on? Can it evaluate its own decision? What happens when it begins to consider the basis of its thoughts? At the beginning of the course we will distinguish between these questions in order to gradually clarify the relationship between them.

David Chalmers suggested that there should be a separate discussion of the explanation of functions—for example, stimulus discrimination and state reporting—and the explanation of why what happens is experienced. This is an influential philosophical proposition, not a pre-established prohibition against explaining experience in terms of physical mechanisms. For our route, it is useful because it requires us to name exactly what the next model explained. [2]

Reflection begins with consideration of one’s own experience, judgments or actions, allowing for their revision. You note: “I considered the weakening of the coolness to be evidence of heating, although I did not measure the temperature.” Now we can ask what other changes have occurred. The analysis continues with observation: hold the position of the hand, clarify the time, compare conditions. This is how reflection becomes part of the research.

Worked example

Here is a teaching note, not the result of a published experiment: “Touched a pen. I felt coolness in my fingers. Later it became weaker. The touch felt slightly pleasant. The handle probably got hot from your hand. I don’t remember at what point I noticed the change.”

The first sentence sets the action. The next three describe the experience. The fifth offers a reason. The latter marks the limitation of recollection. None of this needs to be thrown away: research needs the conditions of the episode, the description, and the ideas that arise. You just need to maintain the differences between them.

The next step is to clarify the word “weaker.” Has the coolness become less intense or have you noticed it less often when looking at the glare? These versions are not yet identified causes. They indicate which additional issues to focus on for the next observation. A good clarification narrows down uncertainty, rather than simply adding a fancy term.

How to parse one recording without losing the experience

Let’s take four sentences: “I touched the handle of the cup. There was a feeling of coolness in my fingers. This cup is comfortable. It probably cooled off by the

window.” The first reports the action, the second describes the quality of the experience, the third expresses an overall assessment, and the fourth offers a reason. We do not choose one correct sentence: each can be useful if its role is preserved.

Now let’s add: “The coolness was pleasant.” Here, pleasantness may denote a directly experienced tone rather than an assessment of the merits of the cup. Therefore, the marks refer to the meaning of the phrase and not to individual words. If in doubt, clarify: are we talking about a current touch or a stable relationship to the object? This is the first example of a research question arising from careful description.

What has now been established

We now have three different items to test: the event, the experienced quality, and the presumed cause. The ability to distinguish between them is more useful than a long list of impressions. Now apply this distinction to our own short record.

Practice 1 First record of experience

Why do it: Learn to keep a description of an experience separate from an assessment of a thing and a version of the cause.

What you will need: No special training required. All you need is an ordinary object at room temperature, a sheet of paper and a few quiet minutes.

Time reference: 15–20 minutes. Can be done intermittently; This is not a speed standard.

1. Look at the object or touch it for about a minute. Immediately after this, write down five sentences. Indicate when the recording was made.
2. Label the descriptions of the experience O, general assessments C, hypotheses about the cause D. Write down actions and external conditions separately. If the sentence is mixed, split it.
3. Choose one vague phrase. Specify location, time, or quality, adding only what you can actually communicate.
4. Repeat the observation. Save both entries and name one difference between them and one question that has not yet been resolved from these entries.

What to Keep: Two dated notes with notes and one specific question for the next observation.

Hint: If you can't select a mark, don't guess. Write down two possible meanings of the sentence. The inability to clarify the content also limits future inference.

Exploratory continuation: Test one version of sensation change, changing only the selected condition. For example, keep the method of touching, but change the order of observations. Before repeating, write down the expected difference.

Analysis and answers - Appendix F, chapter 1, p. 279

Self-test questions

1. Why does the phrase “the object became hot, so the coolness disappears” contain more than a description of an experience?
2. Can pleasantness be considered a characteristic of experience?
3. What is established if the second record contains twice as many details?

Use these notes as a guide in the next chapter. We will move from one episode to several and try to describe the similarity relationships that the future model should preserve.

Chapter 2

Differences and similarities of experiences

The first recording made it possible to separate the quality of an experience from the story about its cause. But defining each sensation separately does not yet show how the differences between sensations are arranged. Now let's collect several episodes and ask the same comparison question. The similarity map will be the first object by which the quality of the model can be judged: it should explain the relationships between cases and predict new comparisons.

From individual experience to relationship

You recorded how it feels to touch the mug. There is a cloth napkin nearby. Touch is clearly different, but the answer “they are different” does not help the research much. You can clarify: one seems smoother, the other warmer; In one person the sensation occurs in a small area of the finger, in another it is distributed more widely. Now the difference breaks down into several relations that can be compared.

This chapter is devoted to the qualitative organization of experience. We will ask in what ways experiences are similar, what differences persist across changing conditions, and how to represent the results so that the model can explain them. The first useful model often begins with the relationships between cases, rather than with a final definition of each case individually.



Figure 3. Comparison of surfaces in one episode. Object illustration of the exercise: Ceramics and fabric offer different touch conditions. The reader compares the qualities noticed and records the intensity and duration separately; the properties of materials are not a substitute for describing the experience.

Quality, intensity and duration

Imagine two sounds of the same volume: a short knock and a long hum. Their difference cannot be exhausted by the word “stronger.” Volume, sound character and duration answer different questions. Likewise, slight roughness does not automatically turn into coolness when its intensity increases. To compare experiences, you need to indicate which relationship is being considered.

We will call quality the way a sensation is experienced. Intensity indicates its severity in the chosen relationship. Duration refers to the length of time. These are working distinctions, not a ready-made list of independent coordinates for any experience. In a particular case, the features may interact: a longer sound may seem more intrusive, and a change in context may affect its noticeability.

It's helpful to ask consistent questions. What has changed? On what basis? Was this change permanent or just at the beginning? If you can't separate the two aspects, write it down as a property of the current description. Research rigor lies in selecting appropriate distinctions rather than mandating completion of a one-size-fits-all questionnaire.

Names and experienced similarities

The same verbal category can unite different impressions. The word "rough" applies to fabric, wood and paper, although they don't feel the same. The opposite situation is also possible: two objects have completely different names, but feel similar in some respects. Therefore, the coincidence of names and the similarity of experiences should be recorded separately.

Let's say you are comparing four surfaces. Two belong to metal objects, two belong to plastic objects. If you ask to distribute objects by material, you will get one grouping. If you ask about the smoothness of the touch, the result may be different. Neither grouping is wrong in itself: they answer different questions.

This is an important difference for the future model. A system that's good at guessing the name of a material doesn't necessarily reproduce the relationships between your touch ratings. You need to decide in advance what behavior the model should predict. Otherwise, success in a convenient task can be quietly mistaken for explaining another phenomenon.

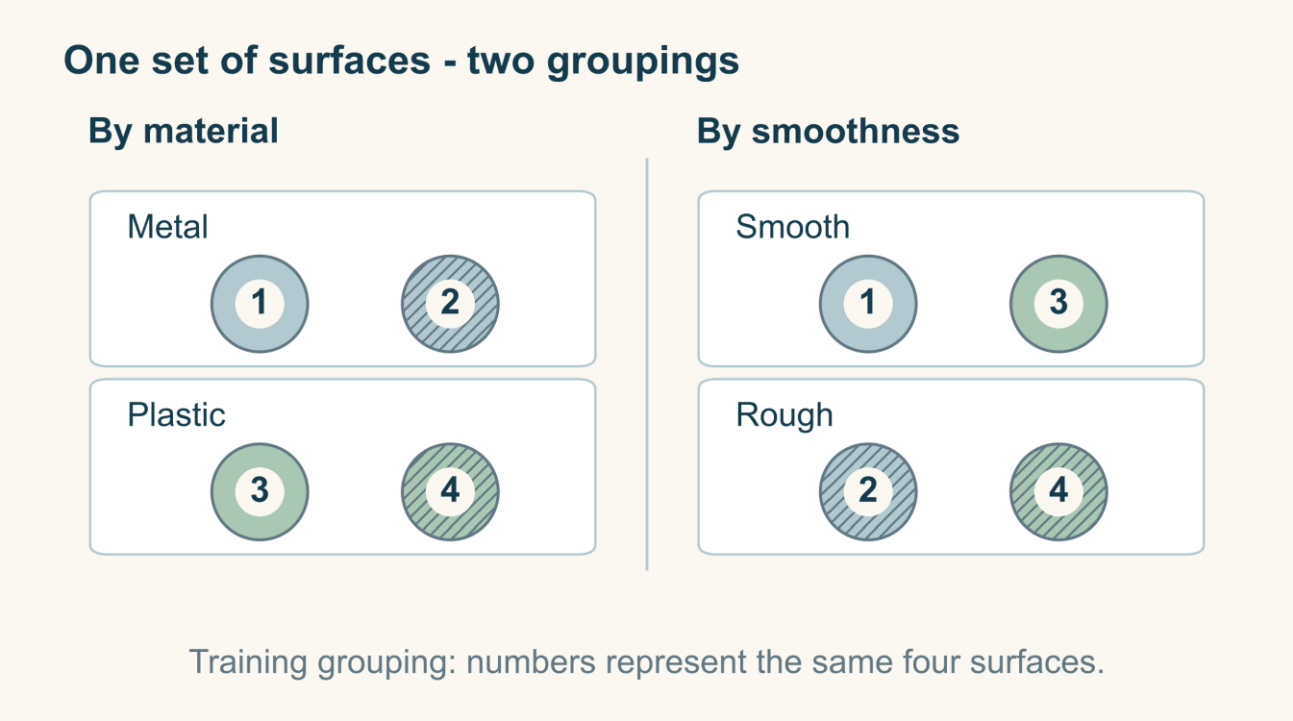

Figure 4. Names of materials and similarities of touch. The training set contains two metal and two plastic surfaces. The same items are grouped by material and by smoothness: the question chosen determines what attitude the model should reproduce.

How to create a similarity map

Let's call the selected episodes A, B, C and D. The letters indicate instances of observation, not properties of experiences. For each pair, you can answer which touch seems more similar to the other, or give a conditional assessment of the

difference. It's easier to start with ranking: which pair is most similar, which is the least, and which are difficult to compare.

With four cases, there will be six different pairs. This already allows you to see the structure. For example, A and B may form a close pair, as can B and D, while the differences between these pairs are more noticeable. This results in a preliminary map of two groups. It is more useful than four independent epithets because it presents models of relationships that should be reproduced.

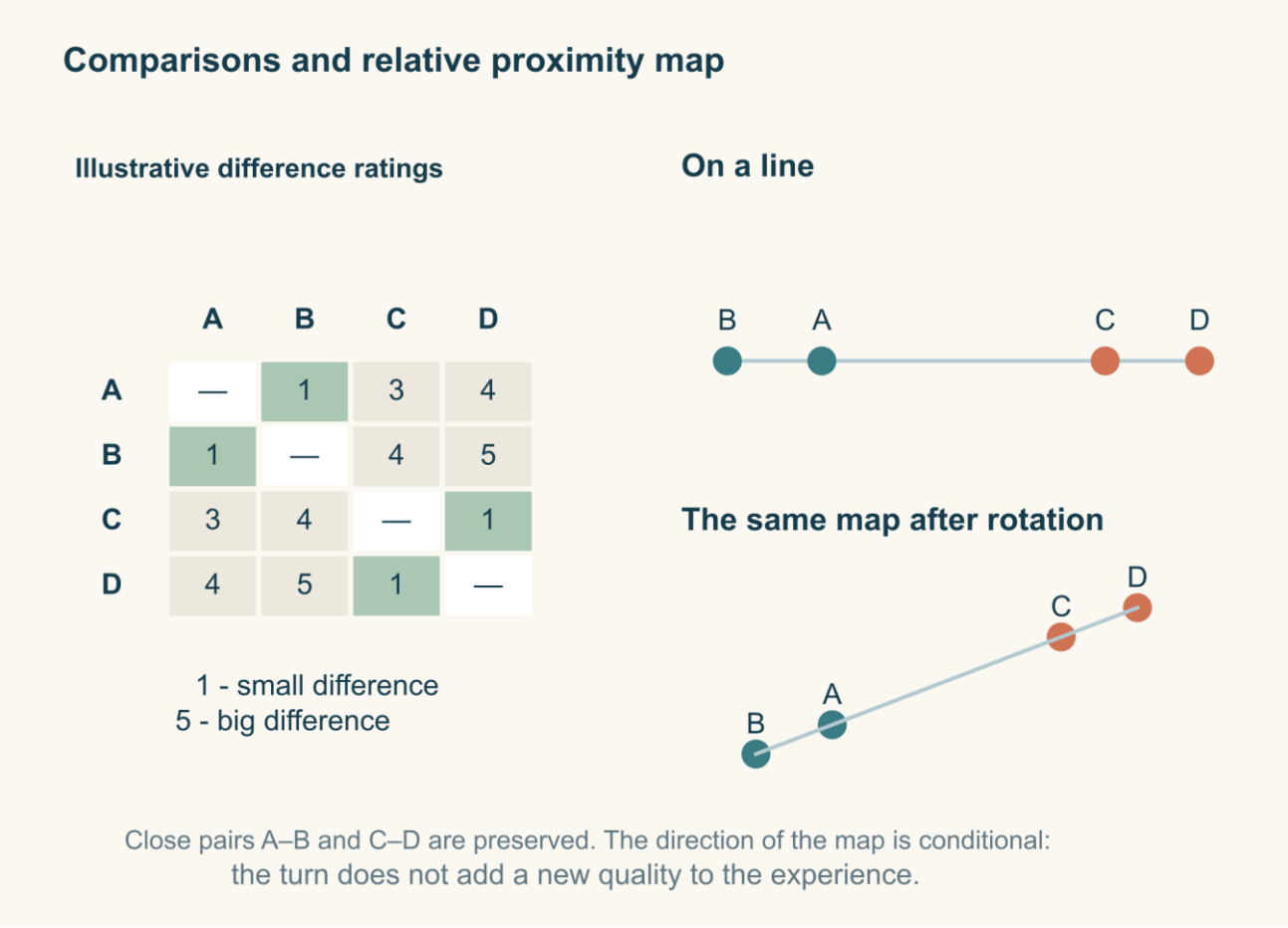

Figure 5. From pairwise estimates to a similarity map. The matrix contains six learning assessments from the chapter. On the line, the points are located in the order B, A, C, D with coordinates 0, 1, 4 and 5; the bottom right shows a rotation of the same configuration. This is a convenient presentation of the example data, and not a basis for considering the scale of experienced difference to be a scale of precise relations.

In the methodology of psychological measurements, Roger Shepard proposed to reconstruct the spatial configuration from data on the proximity of objects when the exact distance function is unknown. This approach helps to explore the multidimensional structure of similarity ratings. Our first practice uses only the general idea of arranging cases according to the relations between them; a full-fledged computational method requires additional prerequisites. [3]

Worked example with training data

Below are made up teaching assessments of the difference. A one means a small difference, a five a big one. These numbers illustrate reasoning and do not describe the results of a human study.

A couple of episodes	Difference Score
A and B	1
A and C	3
A and D	4
B and C	4
B and D	5
C and D	1

The most noticeable pattern is two close pairs: A with B and C with D. However, an additional difference remains: A is somewhat closer to B than to D. If you draw two completely indistinguishable points within the group, some of the information will be lost. The map should show both grouping and differences between individual elements.

Suppose the researcher proposed an explanation through one feature - smoothness. He places A and B side by side, C and D side by side at the other end of the line. To check, you need to compare the remaining pairs: is their order obtained without additional permutations? If not, perhaps a second feature or another form of the model is needed. But it is also possible that some estimates are unstable. These explanations will require different next steps.

The estimate 4 cannot automatically be declared to be twice as much an experienced difference as the estimate 2. The reader could use the numbers only for order. Before calculating the exact distances, you need to find out the meaning of the scale. In initial work, it is sufficient to retain statements such as “this pair differs more than that” and note vague comparisons.

Context becomes part of the task

Imagine that you first examined the surface calmly, and then compared it immediately after another touch. Even if the item itself did not change, the comparison conditions became different. In the second situation, there is a recent

example nearby, in relation to which you select words and evaluation. This is a reason to include the order of presentation in the description of the procedure.

Another context asks the question. The request to evaluate smoothness directs the search to some features; please choose the most pleasant item - over others. If the examiner changes the question between trials, the differences in responses cannot be entirely attributed to changes in sensation. To find out the reasons, you need to save the wording of the task or specifically compare several wordings.

This is where the useful concept of invariant comes into play. This is the name given to a relationship that is preserved under the change in question. If renaming A to K does not change the model's predictions, it does not depend on a random letter. If turning a drawn map preserves all proximity relationships, the direction “north” in the drawing does not carry any psychological meaning. Determining what should be preserved is as important as explaining change.

How to read a map without making unnecessary assumptions

In the training table, the difference between A and B is 1, C and D are also 1, and between B and D is 5. First, it is enough to restore the order: the first two pairs are close, the last one is noticeably different. If you place B, A, C, D on the line at points 0, 1, 4, 5, all six numbers in this particular example will match the distances. But it does not follow from the luck of the image that any subjective scale has the properties of length. Therefore, when working with your answers, first check the order of proximity, and use exact arithmetic relations only after justifying the scale.

What has now been established

We compare attitudes on a selected issue. There is no need yet to find the only correct place for each sensation. It's important to keep your original answers and see where the image matches them and where it adds your own assumptions.

Practice 2 Four episodes and one question

Why do it: Obtain a verifiable similarity map and separate its properties from its legend.

What you will need: Ability to separate description and reason from chapter 1. We need four available surfaces and one constant comparison question.

Time reference: 25–35 minutes. Can be done intermittently; This is not a speed standard.

1. Name the four surfaces A, B, C, D and briefly describe each touch.

2. Write down one question, for example about the similarity of the nature of the touch. Compare exactly six pairs: AB, AC, AD, BC, BD, CD.
3. Order the pairs by difference, leaving equalities and doubts. Draw four points first on the line, then on the plane. Note one attitude that has improved and one that has not been expressed.
4. Repeat the comparisons in a different order with the same question. Save both cards and differentiate between the change in response and the suspected reason for the change.

What to keep: Six paired comparisons, two depictions, and a list of stable and unstable relationships.

Hint: Rotating the drawing does not change the distances. Adding a second coordinate is only useful if it expresses a certain relationship better than the first version.

Exploratory continuation: Add a fifth surface. Before making a new comparison, predict at least two of its similarities according to the proposed explanation, and then check the prediction. Manual rearrangement after each answer does not count as a prediction.

Analysis and answers - Appendix F, chapter 2, p. 279

From map to hypothesis

The resulting map describes the structure of your answers. The next task is to connect it with the mechanism. For example, you can assume that the estimates depend on the combination of two features, and test a new case that was not involved in the construction of the map. If the position of a new case has to be assigned manually each time, the explanatory work has not yet been done.

In what follows we will ask what relationships between the states of an artificial system could correspond to the differences found. The first requirement for such a project is already clear: it must predict the structure of new comparisons. The question of whether its own states are accompanied by experience will require an independent theoretical step.

Self-test questions

1. Why does the coincidence of verbal names not guarantee the same experienced quality?
2. What does it mean to evaluate the difference 4 on a conventional scale from 1 to 5?
3. How can you check whether an additional coordinate helps the explanation and not just decorates the map?

The map described the differences between experiences. Now let's find out how the content relates to one's own body, action and past.

Chapter 3

What It Means to Experience Oneself

We have learned to compare the qualities of experiences. However, two similar impressions may relate differently to the experiencer: the movement of the hand and the movement of the pointer are both visible, but the hand is felt to be part of the body, and the pointer usually remains an external object. To further explore this, let us separate perspective, bodily identity, authorship of action, and self-image over time. This separation will be needed later when building event source models.

Multiple meanings of one word

You reach out to the pencil and say, “I took it.” A short phrase combines several relationships. The pencil was visible from a certain place. The hand felt like part of your body. The movement was perceived as your action. Do you remember why you needed a pencil? This is enough in everyday conversation. To investigate, we have to ask what kind of relationships we want to explain and whether they always necessarily occur together.

Let's start with perspective. The table is visible from here: one side is closer, the other is partially closed. The sound comes from the left of your position. Perspective is the organization of content relative to some point or frame of reference. It is different from biographical knowledge about oneself. To see the edge of the table on the right, you do not need to mentally say your name or remember your profession. However, the presence of coordinates calculated relative to the robot's camera does not in itself establish the experience of perspective. In describing a system, it is necessary to distinguish between the coordinate structure and how the situation is experienced.

The second question concerns the ownership of the body. The feeling “this is my hand” is not the same as a judgment of legal ownership. We are talking about how a part of the body is given in experience. The third question concerns the authorship of the action: “I produce this movement.” The fourth is the image of myself in time: what kind of person I consider myself to be, what I remember about my past, what intentions I associate with the future. These four questions provide a useful work map, although they do not exhaust all uses of the word “I.”

Ownership of the body and authorship of the action

Let's look at a common difference. You deliberately raise your hand, then calmly place it on the table. In the second case, the hand does not cease to be yours due

to lack of movement. Ownership of a body cannot be defined as continuous voluntary control over it. Now you move the mouse: the movement of the pointer on the screen is felt as a result of your action, although the pointer itself is usually not felt as part of the body. Control of an object and body ownership also do not coincide.

In a study by Andreas Kalckert and Henrik Ehrsson, participants observed the movements of an artificial hand in relation to the movements of their own. The authors changed the consistency and nature of the movement, as well as the position of the model. The result was conditions in which assessments of ownership and authorship diverged. The work supports research separation of these components of bodily self-perception; it does not turn each of them into an independent subject. [4]

Another distinction is useful here. A sense of authorship is a characteristic of an experience or an account of it. Actual causal involvement is a characteristic of events. It is possible to press a key, not notice the effect, and still cause a causal effect on the program. You can expect that the movement was performed by you, although at that moment the recording played it. To test causal participation, we need information about the structure of the situation, and to study experienced authorship, we also need data from the participant. One question does not replace another.

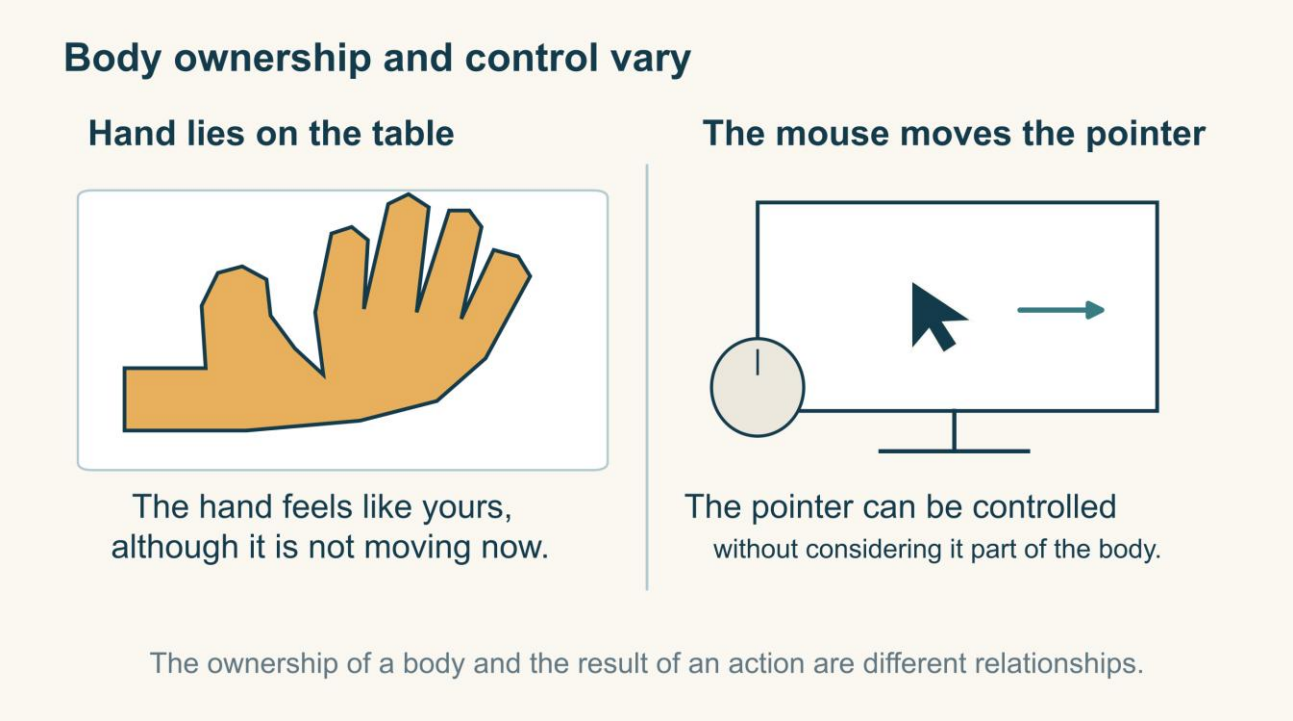

Figure 6. Belonging without movement and controlling an external object. The immobile hand may continue to feel like part of your own body. Pointer control refers to the relationship of action to result, not automatically to the physicality of the pointer.

How the body becomes accessible without looking

Try to describe the position of your hand lying on the table without looking at it specifically. Finger flexion, surface pressure, and approximate elbow position may be available. Information about the position and movement of body parts is called proprioceptive. It participates in bodily orientation along with vision, touch and other sources. The term denotes a specific type of information and sensitivity, rather than a separate small consciousness within the body.

Consistency of sources is important for the researcher. If you see a hand where you expect to feel it, one group of explanations links the information to a single bodily object. If the information differs, the question arises as to what the system gives more weight to and why. For now, this is an outline of the research problem. In chapter 11 we will turn it into a computable model and formulate predictions for different conditions.

Body ownership cannot be reduced in advance to a single signal match. Two lights can flash at the same time without becoming parts of your body. The model must explain which sources are compared, how the shape and position of the object, the history of interactions and possible alternatives are taken into account. It is precisely such clarifications that turn the expressive metaphor “the brain assembles me” into a concrete task.

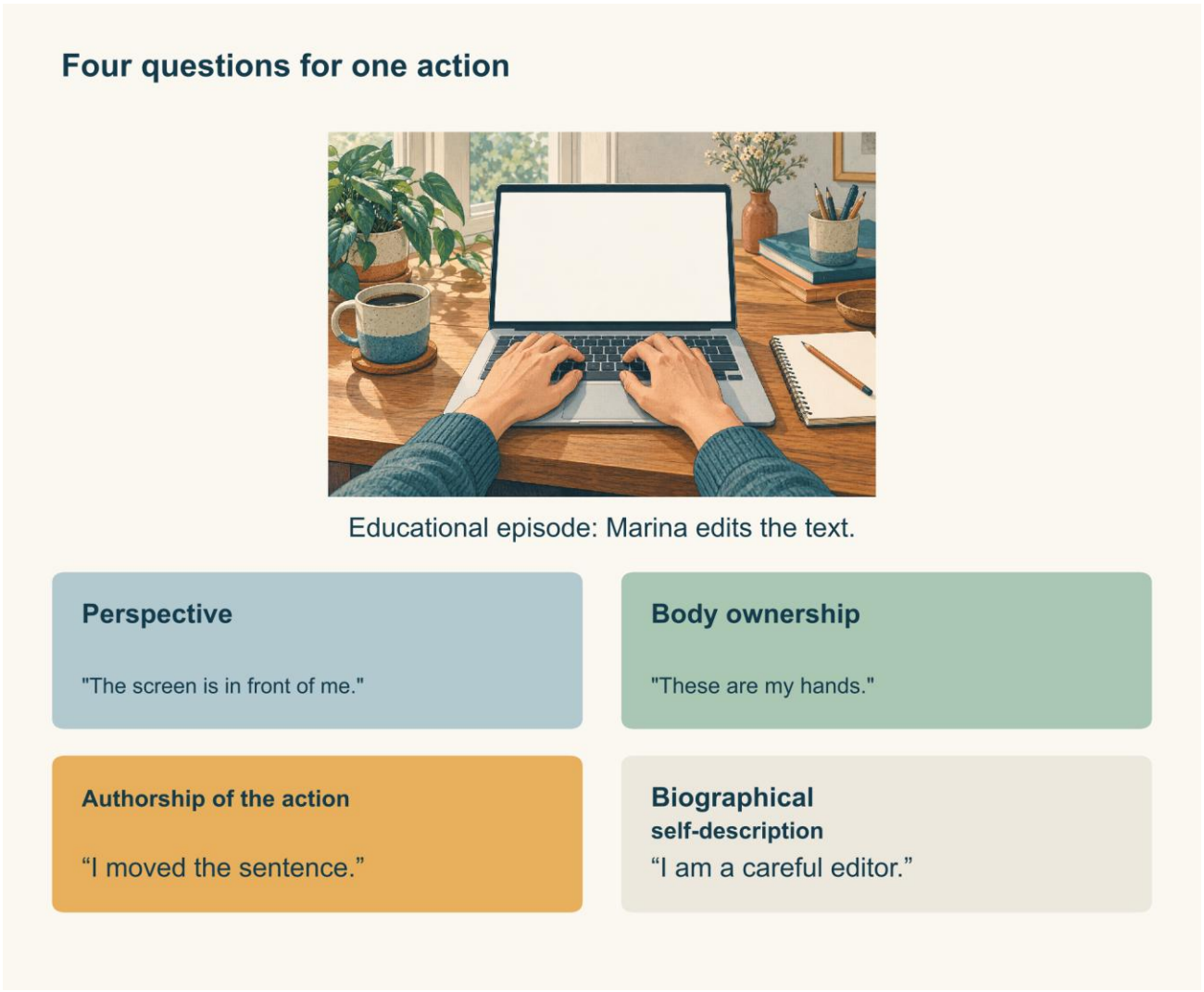

Figure 7. Four attitudes towards yourself in one episode. Perspective, bodily identity, authorship, and biographical self-description are highlighted separately. Clarifying whether a person or a program corrected a gap refers to the authorship of a particular change, and does not automatically apply to all other relationships.

An example with a text editor

Let's imagine a training episode. Marina transfers the sentence in the document. She moves the mouse, sees the highlighted text, and intends to move it to the beginning of the paragraph. After releasing the button, the program automatically corrects the space. Marina says: “I moved the sentence and corrected the format.” Let us consider this phrase without assuming that it necessarily accurately describes all events.

Moving the sentence was Marina's intention and action. Correcting the space could be an automatic operation of the program. The final result is visible to her: she can check the arrangement of words. But access to a result does not mean access to all the reasons for its occurrence. In the research record, it would be worthwhile to

separately indicate the movement of the mouse, the change in the document, Marina’s message, and the automatic correction rule known to the researcher.

Let’s say that after viewing the history of changes, Marina clarifies: “The program corrected the gap.” This is a revision of the judgment of authorship. It does not follow from this that the original sense of control was invented. It could arise from the connection between intention, action and overall result. The research challenge is to explain this relationship, including when it leads to overly broad self-attribution of the outcome.

The biographical self-description appears on another level: “I am a careful editor.” Autospace alone is not sufficient to verify this statement. It generalizes to many situations and depends on the criterion of attentiveness. If the levels are mixed, a small error in determining the cause becomes an unwarranted inference about the individual. When recording separately, there remains a specific question that can be checked.

Is an internal observer necessary?

Self-awareness is sometimes thought of as a spectator looking at an internal screen. This picture is convenient until it becomes an explanation. If the viewer must recognize the image, remember the past and understand that it relates to him, a smaller copy of it has appeared inside the original task. It will be necessary to explain the abilities of this viewer.

A more productive initial question is what information is associated with the system and how this connection is used. For example, a program can store its own identifier, distinguish its commands from incoming messages, and evaluate the reliability of its response. These are different functionalities. They can be implemented and tested individually. The presence of an identifier does not prove an experienced self, but it does provide a clear object for the study of self-reference.

Explicit reflection adds another layer: the system or person views its own judgment as subject to verification. The phrase “I can see the edge of the table” and the question “Why am I sure that I can see it accurately” serve different purposes. When constructing hypotheses about a minimal subject, one cannot imperceptibly make complex verbal reflection a mandatory condition of any experience. Such a condition would require a separate justification.

Why Control Doesn't Equal Ownership

In the first episode, the hand rests calmly on the table. There is no action now, but her position and touch can be experienced as relating to her own body. In the

second episode, you move your mouse and see a pointer. Its movement depends on the command, but this does not imply a physical identity. In the third, you read yesterday's note: its creation relates to your history, although the current reading is not the experience of the current writing. These cases show why separate description columns are needed. One general answer, “It’s me,” would hide the differences.

What has now been established

Four questions refer to different relationships within the same episode. There is no need to attribute independence to them in all situations. First, it is enough to learn to notice what answers are available and what attitude is being explored.

Practice 3 Four questions for three episodes

Why do it: Develop several meanings of self-reference in three ordinary episodes.

What you will need: Distinguishing between an event and a report from chapter 1. Remember that actual causal involvement and sense of authorship answer different questions.

Time reference: 20–25 minutes. Can be done intermittently; This is not a speed standard.

1. Describe three cases: a stationary hand on a table, controlling a pointer, and reading one's own previous text.
2. For each, separately indicate the perspective, the experienced ownership of the body, the authorship of the current action, and information about yourself in the past.
3. Choose one causal claim that can be independently verified: what exactly caused the movement or when the text was created.
4. Match this check with the first description. Save the discrepancy if there is one; do not turn it into a general assessment of personality.

What to Keep: Three descriptions with four separate questions and one independently verified relationship.

Hint: The absence of words does not prove the absence of a component. You can write down exactly which question remains unclear and what clarification would help answer.

Exploratory follow-up: Suggest a change to the situation that could separate authorship and ownership. Formulate a separate forecast for each indicator and explain what external events need to be controlled.

Analysis and answers - Appendix F, chapter 3, p. 280

Self-test questions

1. Why doesn't controlling a pointer on a screen prove the experience of the pointer as a part of the body?
2. How does actual authorship of an action differ from the feeling of authorship?
3. Why is knowledge of one's own name not sufficient for a complete model of the experienced self?

Having distinguished attitudes towards ourselves, let’s move on to assessing our own thinking. The next chapter will ask how much the feeling of understanding allows us to judge the correctness of a decision and how to check this connection.

Chapter 4

How we notice our own thinking

The sense of self includes the ability to notice one's own decision, but the noticed decision still needs to be evaluated. The familiar feeling of “everything is clear” sometimes precedes the check. In this chapter we will distinguish between the correctness of the answer, confidence before feedback, and subsequent explanation. Maintaining their order will allow you to see what exactly has changed after the check, and turn reflection into the basis of a new action.

Thought and knowledge about thought

Read the phrase: “The box contains fewer pencils than pens, but more pencils than erasers.” You can feel that everything is clear and immediately move on. You can also stop and ask which items are in abundance. In the second case, the initial impression is verified by action. If you have to read the condition again to answer, this does not necessarily mean a lack of understanding. However, it becomes clear that the feeling of understanding and the willingness to complete a certain task are different data.

In this book, we refer to introspection as referring to one's own current or recently experienced state. Reflection means considering one's own thinking, action or judgment as a subject for analysis and possible revision. Metacognition is a broader family of processes for assessing and regulating cognition. For example, the choice to retest a difficult answer may depend on an assessment of its reliability. These operational definitions help to ask different questions; in scientific literature, the boundaries of terms may be drawn differently.

Having a thought does not guarantee detailed knowledge of how it came about. Let's say you quickly recognized a familiar word. The recognition result is available to you, but you may not know the sequence of letter processing. The story “I looked at each letter first” then requires verification. It may be a description of actually noticed steps or a plausible reconstruction after recognition.

Why Sincerity Doesn't Bring Accuracy

A self-report may be completely sincere and yet inaccurate. The reason is not necessarily a bad memory: sometimes it is not clear what differences should be noticed and what words to express them. The question “was the thought verbal” requires distinguishing the spoken internal phrase from knowledge of its meaning and from the phrase that appeared during the answer itself.

Eric Schwitzgebel, in a philosophical article on the unreliability of naive introspection, examines the reasons to doubt the infallible knowledge of one's own current experience. His examples concern, among other things, visual clarity, emotions, and thinking patterns. This is an argument against the automatic infallibility of self-reporting, not a measurement of the error rate of any individual, nor a reason to reject self-reporting. [5]

For practice, a modest but effective rule follows from this: separate the moment of experience, the moment of the first message and the moment of subsequent explanation. If, when solving a problem, the answer appeared first, and the verbal justification appeared later, write it down. There is no need to fill the gap with supposed mental steps. A gap in the record is more useful than fictional continuity.

Confidence must relate to a specific issue

The phrase “I am eighty percent sure” becomes research useful after clarifying what exactly the certainty is. Is the answer correct? Exactly memories? The ability to repeat the explanation tomorrow? These are different events, and the way they are verified is different. A number without specifying an event creates the appearance of precision.

In problems with a verifiable answer, it can be agreed that confidence refers to the probability of the correctness of a given answer. If a person chooses the 80 percent level many times, it is desirable that approximately eight out of ten such answers turn out to be correct. This is the idea of calibration. One correct or incorrect answer does not allow us to evaluate it: the probabilistic expectation is checked against a series of comparable cases.

Stephen Fleming and Hakwan Lau distinguish between the tendency to give high confidence, the ability to confidently separate correct answers from incorrect ones, and the effectiveness of such assessment in relation to the success of the main task. These are different measurement questions. High average confidence by itself does not show good discrimination of one's own errors. [6]

Save the task, the answer, the confidence before the test and the result of the test. A complex calculation will not restore the lost original forecast: the confidence recorded after reviewing the solution answers a different question.

An example of two readers

Let's present the training data created for this chapter. Anton and Vera solved four identical problems each. Both performed the first and third correctly, but made mistakes in the second and fourth. Their success in this small series is the same. Confidence was recorded before feedback was received.

Quest	Correctness	Anton's confidence	Vera's Confidence
First	That's right	90 percent	90 percent
Second	Error	90 percent	40 percent
Third	That's right	90 percent	80 percent
Fourth	Error	90 percent	30 percent

Anton assigns the same confidence to every answer. In this series, Vera places less trust precisely in the answers that turned out to be wrong. Equal numbers of correct answers therefore do not imply equal quality of self-checking. If only two tasks may be checked again, Vera's ratings help direct effort where it is needed most. This is one possible practical function of assessing one's own thinking.

The four tasks do not establish a stable property of a person: random coincidence and different understanding of the scale are possible. To check, additional tasks and the same instructions are required. The table describes the quality of the response assessment and does not allow Vera to be declared “more conscious”.

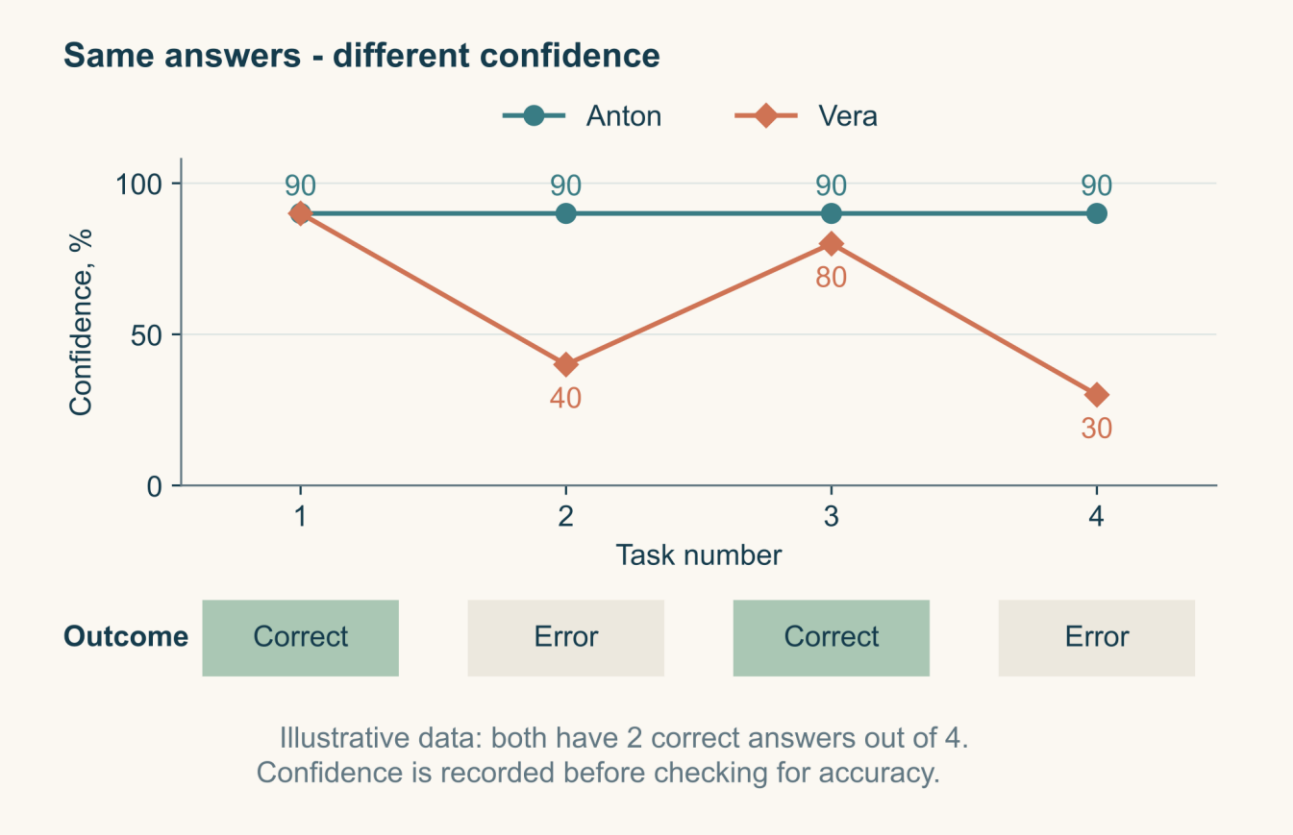

Figure 8. Confidence with equal success. In the educational series, Anton and Vera correctly answer tasks 1 and 3. Anton attributes 90 percent to all answers, and Vera attributes 90, 40, 80 and 30 percent. On these four occasions, Vera's assessments highlight future errors; new data are needed to draw conclusions about sustainable capacity.

From self-observation to changing the way of solving

Reflection produces results when it changes the testable decision about the next step. Suppose you notice, “When reading comparisons, I often keep the first pair of objects and lose the third.” From observation, a proposal arises to write down relationships with arrows. You can then check whether errors have decreased on new tasks of the same type. This creates a small research chain: difficulty, suspected cause, change in procedure, new results.

However, successful change does not prove the whole story about the cause. The arrows could help simply by increasing the operating time. To distinguish the explanations, you will need a condition with the same time, but without the arrows. We will examine this issue in detail in the chapters on verification. What is important now is to see how reflection moves from a general “we need to be more careful” to a specific change for which competing explanations are possible.

Self-control can occur without extensive introspection. The person notices the uncertainty and rereads the line without constructing a theory of his reading. Explicit reflection begins where the basis of an assessment, a method of action, or

the boundaries of a conclusion are considered. It is useful to consider automatic error correction and verbal investigation of the cause separately.

How to keep knowledge in order

Let the participant write down the answer and confidence of 90 percent, then see the error and say that he doubted it from the very beginning. The last message may be sincere, but it comes after new information. Both records are kept for the study: the initial prediction and the subsequent interpretation. If you replace the first recording with a second one, the ability to assess the initial accuracy of self-control will disappear. Therefore, in the exercise below, the answers are separated from the conditions and placed in a discussion at the end of the book.

What has now been established

We do not require infallible introspection. It is enough to make the moment and the subject of confidence explicit. The subsequent error becomes material for changing the procedure if the original record remains available.

Practice 4 Forecast before inspection

Why do it: Compare preliminary confidence with the outcome and choose a specific method for the next test.

What you will need: Distinguish between the moment of experience, the first recording and the later explanation. Paper is enough for the exercise; knowledge of formal logic is not required.

Time reference: 15–20 minutes. Can be done intermittently; This is not a speed standard.

1. Without opening the analysis, solve three tasks. There are fewer pencils in a box than pens, but more than erasers: what is the least of? All green cards are round, some round ones are large: should there be a large green one? Nina arrived before Oleg, Oleg before Peter: who is last?
2. For each answer, write down the confidence from 0 to 100 percent and one phrase about the observed progress of the solution.
3. After all three entries, open the practice analysis 4. Separately note the correctness and what has changed in understanding.
4. For one difficulty, choose the following technique, for example, finding a counterexample. Don't assess sustainable ability across three tasks.

What to keep: Three pre-confidence answers, test outcomes, and one specific rule for next action.

Hint: The word “some” refers to some of the round cards. Try to imagine several acceptable layouts without revealing the analysis yet.

Exploratory continuation: Test the chosen technique on new similar problems without changing the initial confidence. Compare the accuracy of answers and the ability to separate doubtful cases.

Analysis and answers - Appendix F, chapter 4, p. 280

Self-test questions

1. Why shouldn't explaining a solution after viewing an answer automatically be considered a record of original thinking?
2. Can a person often answer correctly and have poor confidence in recognizing his own mistakes?
3. How can you turn “I don’t read carefully” into the beginning of a research paper?

Now we have an example of a useful revision. To explain why it helps, we need to distinguish the name of the result from the mechanism at work. This is what the next chapter will deal with.

Part 2

Transition to scientific research

Chapter 5

What counts as an explanation?

In the previous chapter, it was possible to improve the solution, for example, start looking for a counterexample. But the improvement still leaves the question of the reason: did a new technique, extra time, or familiarity with the task help? Now let's learn how to construct an explanation so that it can be compared with another version. This will prepare both the experimental part and later models of consciousness.

From title to research question

The man reread the sentence several times and finally understood it. Why did this happen? The answer “understanding turned on” repeats the name of the result. The answer “the brain processed the information” adds to the overall process, but does not yet show what changed between the unsuccessful and successful reading. A good beginning of an explanation clarifies the problem: what connection was made, what information was previously available, and what operations could lead to a new result.

The description tells what happened. For example: “On the second reading, the participant correctly ordered three events in time.” The explanation seeks to show how this result was achieved. A prediction indicates what to expect under certain conditions. These tasks are related, but not identical. It is possible to accurately predict the output of a device from a table of past occurrences, without knowing almost anything about the internals. It is possible to propose a plausible mechanism that does not yet predict real data well.

For research on experience, it is also important to clarify the result itself. Are we explaining a correct answer, a message of clarity, the ability to use information in a new task, or an experienced understanding? One experiment may contain all of these data, but moving from one to the other requires an argument. Otherwise, the theory quietly explains a convenient indicator instead of the original phenomenon.

What do we call a mechanism?

Let's imagine an automatic lamp. The sensor measures the illumination, the program compares the number with the threshold, the switch supplies power. When it gets darker, the lamp turns on. The parts, their interactions and the sequence of changes are named here. If the sensor is covered, the result will change in a predictable way. This is a simple mechanistic explanation that can be tested by intervention.

Peter Machamer, Lindley Darden and Carl Craver proposed to analyze mechanisms through organized entities and activities that produce change. The research process itself is useful for our work: listing the components is not enough, we need to show their organization and operation. Their article presents a philosophical analysis of explanation rather than an experimental theory of consciousness. [7]

Let's take a reading approach. Components of the model may be word representations, a previous sentence store, and a temporal relation matching rule. The words “storage” and “rule” are still relative. It is necessary to clarify what is stored, for how long, how it is used and under what conditions it is lost. Only then will it become clear why the model copes with one text and makes mistakes in another.

The mechanism can be described at different levels. The electrical circuit explains the switching of the lamp through current, and the program through the condition and command. If both descriptions are correct, they may refer to the same system and answer different questions. The mere diversity of levels does not require the existence of several independent causal worlds.

Correlation and Intervention

Let's say that in a training observation, people who looked at the image longer were more likely to say that they understood it well. Several explanations are possible. The extra time helped me figure it out. Complex images both made people look longer and elicited a more cautious response. Or interested participants both looked longer and reported greater clarity. The relationship between two indicators does not yet determine the direction of the causal influence.

The intervention changes the condition according to the research rule. For example, the show time is set in advance and does not depend on the participant’s interest. If other significant differences are taken into account, the comparison becomes more informative about the effect of time. However, the intervention also needs to be dismantled. Longer exposure may simultaneously alter the number of eye movements, rereading ability, and fatigue. These are parts of a possible causal chain or concomitant changes, not automatically resolved issues.

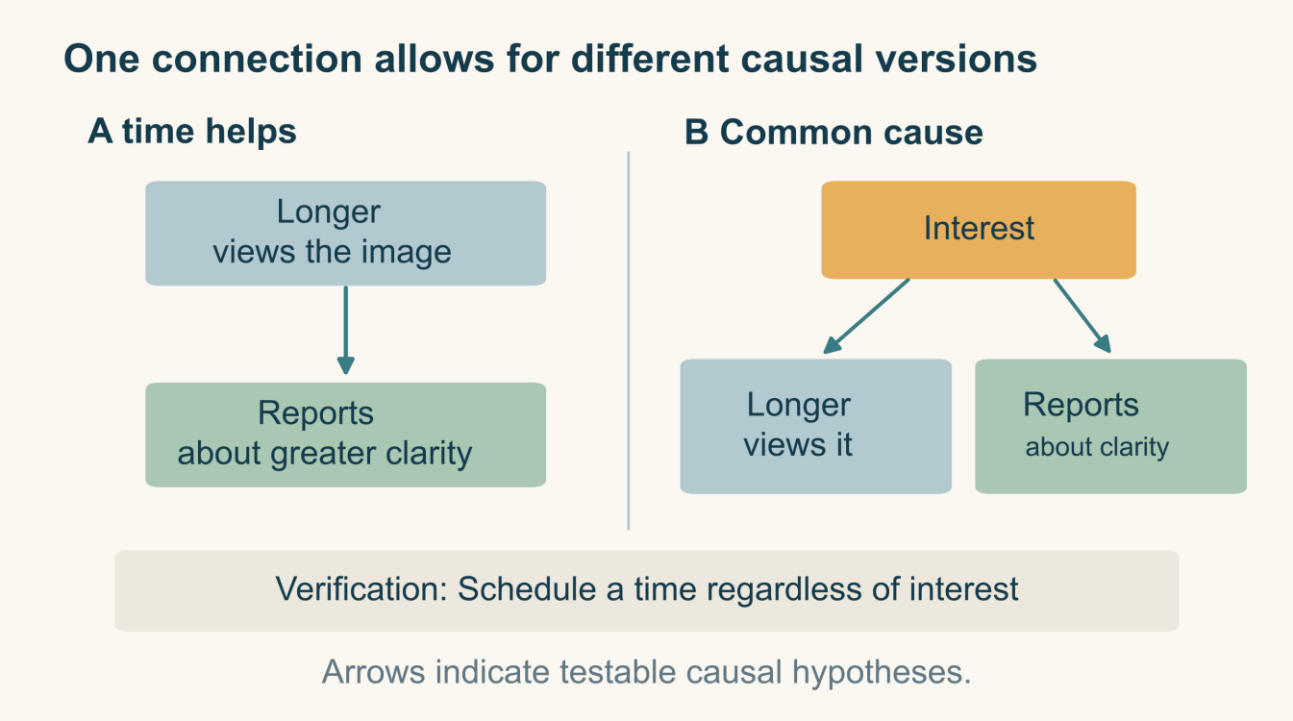

Figure 10. Causal influence and common cause. The relationship between viewing duration and clarity messages suggests different mechanisms: the effect of time or the effect of interest on both indicators. Assigning time according to the research rule helps to distinguish between these versions if other essential conditions are taken into account.

It is useful to distinguish between the cause of an event of interest and an indicator of that cause. The thermometer reports the temperature, but changing the number on the drawn thermometer will not warm the room. Likewise, the neural measure associated with a report may be a cause, an effect, or a contributing process. To choose an explanation requires conditions under which these options make different predictions.

Necessity Sufficiency and the Scope of a Claim

If a property is required for some result, without it the result does not occur in the specified condition area. If the property is sufficient, its presence ensures the result under specified conditions. These relationships cannot be reversed. A working power source is necessary for our lamp to operate, but power alone is not enough: the lamp may be turned off by software or have a faulty switch.

When discussing consciousness, the domain of conditions is especially important. The phrase “process P is necessary for experience” sounds universal. Data from a single visual task in adults support a narrower claim. Generalization to other types of experience, organisms or artificial systems will require additional justification. Refining the scope does not weaken a good theory: it makes visible the boundary between the work already done and the next study.

We must also distinguish causal provision from presumptive identity. A theory can claim that a certain organization of a process is an experience. Then the author proposes a relationship between two ways of describing one phenomenon. Observing co-occurrence does not by itself establish such an identity. The correspondences, limits, and advantages of this view over alternatives will need to be explained.

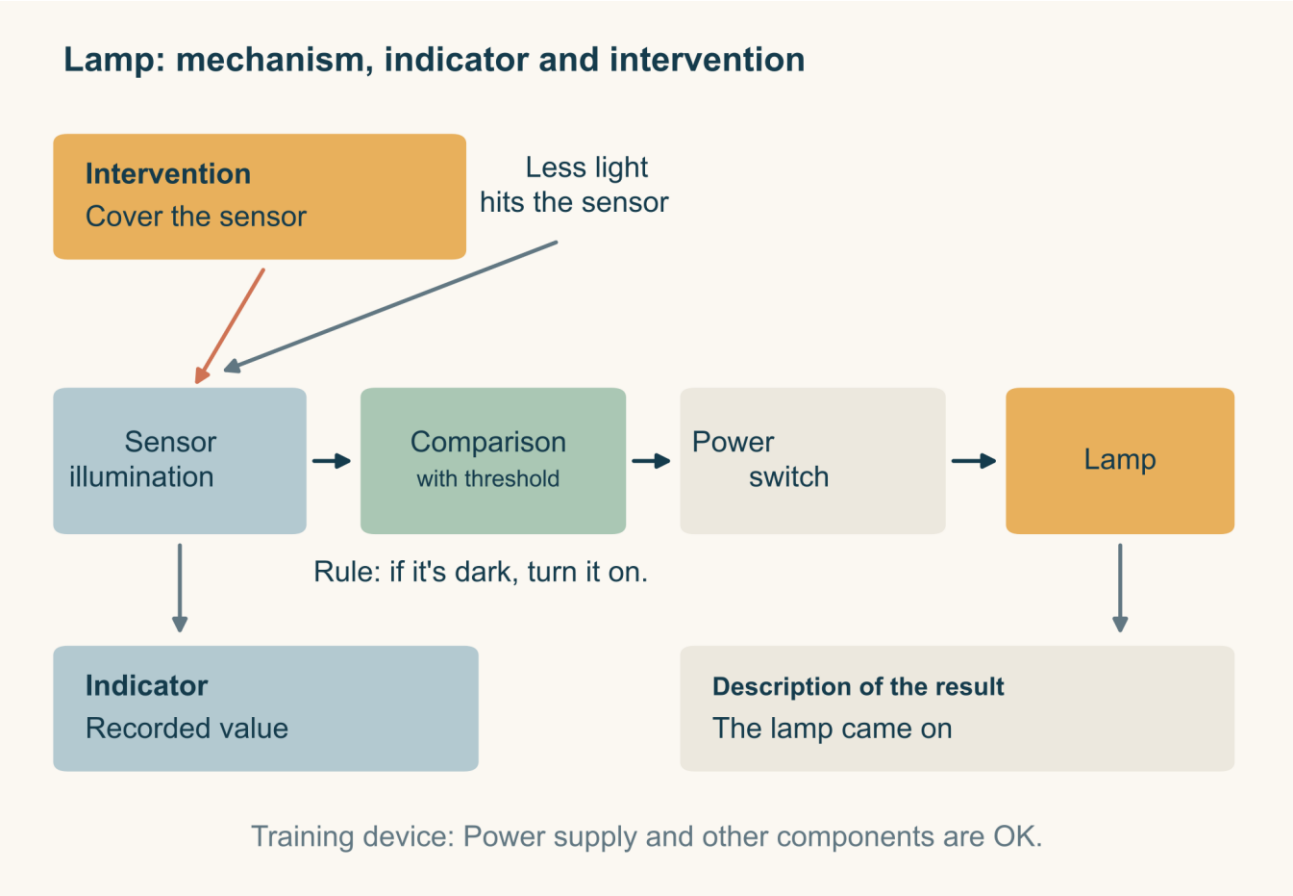

Figure 11. Three research roles in the lamp design. The middle row shows the transmission of measurement, command and power in the training automatic lamp. Closing the sensor is an intervention; the number in the record is an indicator; turning on the lamp - the described result. The forecast depends on the agreed serviceability of the remaining components.

A Worked Argument about Recurrent Processing

Consider the fictional reasoning: “After repeated viewing, people become better at recognizing an image. This means that repeated processing creates consciousness. Let’s add feedback to the program and get an experience.” The first step is an empirical statement about the success of the task. The second connects success with consciousness. The third transfers the proposed mechanism between different systems. Each step adds premises that are not present in the original observation.

Let's restructure the argument. First, we define the indicator: the proportion of correctly recognized images and a separate assessment of visual clarity. Then we

propose a mechanism: the first pass selects the features, the subsequent ones refine them taking into account neighboring elements. The next prediction is that reprocessing should be particularly helpful in conditions where local cues are ambiguous and context is informative.

Now a competitor appears: the improvement simply provides additional computational time, regardless of feedback. Comparison requires models with comparable resources and a task on which their predictions diverge. If the feedback model wins, we will have reason to prefer this mechanism for the chosen outcome. Further treatment of the experience will depend on the individual theory and other tests.

Such an analysis does not close the question of creating a possible subject. It makes the project sequential: first figure out which organization explains what, then formulate a reasoned transition to the properties of experience. Engineering success becomes part of the argument rather than a silent substitute for it.

Designing a Discriminating Comparison between Two Accounts

Let's consider the benefits of a diary. Version one: the saved text holds the details and allows them to be compared. Version two: the pause itself for careful consideration helps. The diary comparison with the absence of any stop is compatible with both versions. A more meaningful comparison gives the same time, but in one condition it saves the record for later use, and in the other it leaves only a pause. Now, the difference in results may limit the explanation. However, we still need to choose a measure of understanding: the length of the story and the accuracy of the new discrimination are not the same result.

What has now been established

The explanation became more definitive as parts, interactions, and various predictions emerged. These requirements do not make the idea true in advance. They allow you to understand what results will really change your attitude towards it.

Practice 5 Restore missing transitions

Why do it: Convert a general statement about the benefits of journaling into a testable pair of mechanisms.

What you will need: The concepts of event, indicator and preliminary forecast. We will distinguish between what is being measured and what mechanism is being proposed to explain it.

Time reference: 25–35 minutes. Can be done intermittently; This is not a speed standard.

1. Select one outcome: for example, correctly distinguishing two previously described types of episodes using new examples.
2. Record two versions: saving the available recording and the extra pause effect. For each, specify the input, action, and expected change.
3. Make a comparison using the same time and material. Describe exactly how the conditions differ.
4. Formulate three outcomes in advance: recording advantage, agreement, and disadvantage. For each, write down the valid output and the remaining alternative.

What to keep: One page with the indicator, two mechanisms, conditions being compared, and interpretation rules.

Hint: If either outcome supports your favorite explanation, clarify the mechanism's promise. The absence of difference due to inaccurate measurement does not yet establish equality.

Research continuation: Add interference to access to stored text after recording. Explain what new conclusion it will allow and what collateral changes will need to be taken into account.

Analysis and answers - Appendix F, chapter 5, p. 281

Self-test questions

1. How does describing a mechanism differ from listing the parts of a system?
2. Why does the joint change in the neural indicator and the report not yet determine the causal direction?
3. What is missing in the transition from a successful recognition program to a statement of created experience?

The next chapter will look at how a selected phenomenon becomes data. For each button, verbal scale, and physiological signal, you will need to explain the relationship to the question under study.

Chapter 6

How to Explore Subjective Experience

We have mechanisms and predictions. Now we need to decide which observations will allow us to verify them. The experience is not transferred directly to the table: a record of the answer is created under certain conditions. Let's consider what the participant's message, the success of the action and the physiological indicator provide. The main skill of this chapter is to reconstruct the path from the question to the data and from the data to a valid conclusion.

Experience does not produce a finished table

You look at the drawing, notice a familiar shape, and press the button. The researcher can save the image, the moment you pressed, your response, and a recording of the physiological signal. The episode of experience itself does not become another file that can be opened next to the others. A bridge is required between the research question and the data: why is this particular answer or indicator informative about what we are studying?

For a beginner, it is useful to distinguish three lines of data. The first is reports about the experience: verbal descriptions, category selection, clarity assessment. The second is behavior in the task: correctness of discrimination, response time, transfer of knowledge to a new case. The third is physiological measurements. They complement each other if the researcher understands how each line is intended to relate to the original question.

Such work is not limited to searching for the most objective instrument. The number also requires interpretation. If the participant pressed the “clear” button, the press in response to the instruction was directly recorded. The conclusion about the clarity of the experience is based on the participant's understanding of the scale, the reporting conditions, and the reporting model adopted by the researcher. This does not invalidate the data: it shows what exactly should be checked.



Figure 12. Participant and observer in the study. Subject illustration of a conditional task with a screen and a response button. The researcher records the activity and obtains a description of the participant; the relationship of this information to experience requires a separate justification in the procedure.

Get a description of a specific episode

Compare the questions “how do you usually perceive music” and “what did you notice at the moment of the last sound.” The first invites generalization and a story about habits. The second directs to a specific episode. If we want consistency in an experience, it is useful to anchor the description to the event and time, and then separate what is directly observed from conjectures about causes.

Claire Petitmengin described an interview method aimed at in-depth exploration of specific subjective experiences. The article discusses the difficulties of awareness and description, ways of working with them and criteria for the consistency of the resulting descriptions. This is a methodological suggestion for data collection and is not a guarantee that any detailed account will be error-free. [8]

For your own practice, ask questions that leave room for an unexpected answer. “What changed before the click?” is usually more helpful than “you felt confident, right?” The second option names the expected experience in advance. Even a negative answer is then built around the proposed category. Once the description is free, specific differences can be clarified while maintaining the participant's original wording.

If a person answers “I don’t remember,” this is also the result of the procedure. You can’t automatically replace it with “I didn’t experience anything.” Perhaps the episode has not been preserved, the question is unclear, or the necessary difference was not noticed. For the next study, you can reduce the reporting delay or change the question. But you can't retroactively get a missing message from a convenient guess.

What Behavioral Evidence Tests

Let's imagine the task: determine which of two lines is longer. Correct selection indicates successful use of information for a given discrimination. If the participant additionally says “the difference was very clear,” a separate report arises. When accuracy and clarity diverge, one metric should not be thrown away for the sake of a consistent story. The conditions for the divergence need to be explored.

The response time is also ambiguous. Slow response may occur due to difficult discrimination, careful checking, or awkward button placement. A quick answer can be the result of an easy decision or a hasty choice. Therefore, time is interpreted together with the task structure and other data. The question “the faster, the more conscious?” doesn't get a reasonable answer without a theory that explains why the connection was expected.

The behavioral task is especially useful when testing access to content. If the information can be used for new output, instruction, or selection, we have obtained information about its functionality. However, access varies by task: a person may correctly indicate the location of an object and find it difficult to describe it in words. One poor response format does not cover all the ways information can be used.

How to read a physiological indicator

Electroencephalography records electrical potential differences through electrodes; conventional functional magnetic resonance imaging uses a signal related to blood flow and blood oxygenation. These are different measurements with different temporal and spatial properties. A colored region in a brain image is the result of processing measurements, rather than a photograph of an individual experience.

Nikos Logothetis and colleagues correlated functional magnetic resonance recordings with electrical neural signals in the visual cortex of monkeys. Their work examined how the measured signal was related to neural activity. For our topic, the step itself is important: the relationship between the instrumental indicator and the process that they want to evaluate is an independent research question. [9]

Suppose the indicator increases before the answer “I see”. It may reflect image processing, message preparation, attention to task, or a combination of processes. To clarify the role, you can change the method of answer, the timing of the question, or the requirement to report the result. Such comparisons are useful when other essential conditions are preserved. The absence of an immediate report removes some of the associated action, but does not yet create a direct measure of experience.

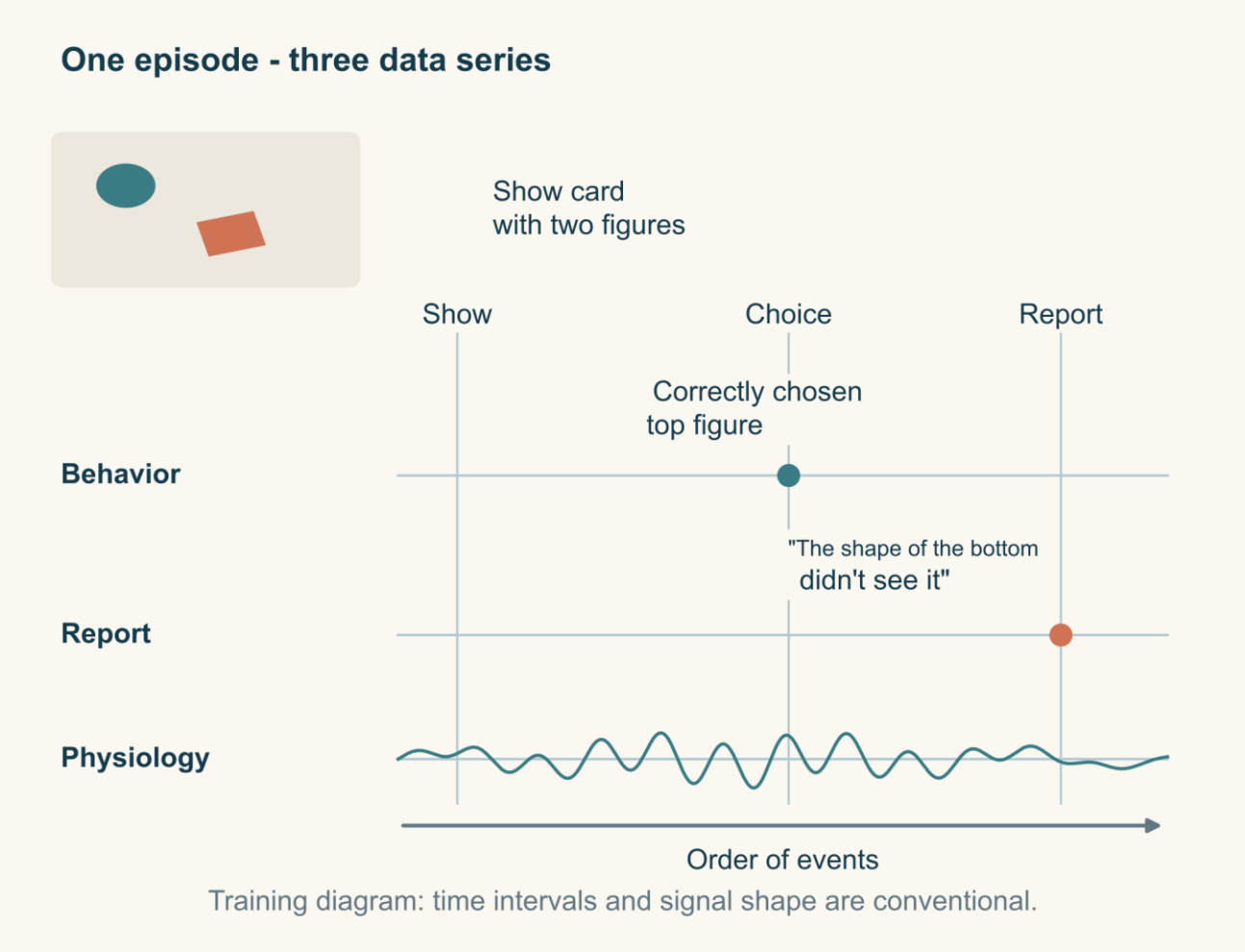

Figure 13. Coordination of different data over time. A fictional episode from the chapter: correct discrimination of the position of a figure is combined with a report of an unclear shape. Pressure, message and physiological recording are represented on a common time axis. The curve shape and time intervals are schematic and do not represent experimental data.

Worked example of mismatched data

Consider a fictitious learning situation. The card shows two figures. The participant correctly chooses which one is higher, but says: “I didn’t see the shape of the bottom one.” There is no necessary contradiction here. Position and shape are different properties. A successful solution to the first question does not

establish a clear perception of the second. The researcher’s error would arise if he designated any correctness with one general label “fully realized the picture.”

Let’s say that the physiological indicator turned out to be higher precisely in cases of correct choice. At least two versions are possible: the indicator is associated with the quality of spatial processing or with confidence in preparing an answer. To distinguish, you can select cases with the same correctness, but different confidence, and separately change the method of reporting. The following study is designed from a specific prediction discrepancy.

Combining methods is sometimes called triangulation. Its strength lies in limiting alternatives, not in voting three devices. If the report, the press, and the neural index depend on a single response preparation, their agreement is not three independent confirmations of the experience. It is necessary to find out which reasons for the indicators are common and which differ.

Why doesn't the correct answer correct the previous description?

After briefly viewing the pen, the person reports its color, but does not mention the inscription. Repeated inspection shows that there really was an inscription. It sets a property of the item, but does not choose between versions: the inscription was not noticed, was noticed and forgotten, or was omitted in the short answer. Another comparison is needed, such as a change in report latency. If the question about the inscription is known before viewing, he directs the search; such viewing cannot be considered the same situation of free observation.

What has now been established

Some data limit the properties of an object, others characterize the use of information and communication about it. The discrepancy between these series helps to pose the question more precisely. Now let's try to describe the procedure so that its meaning is visible to another person.

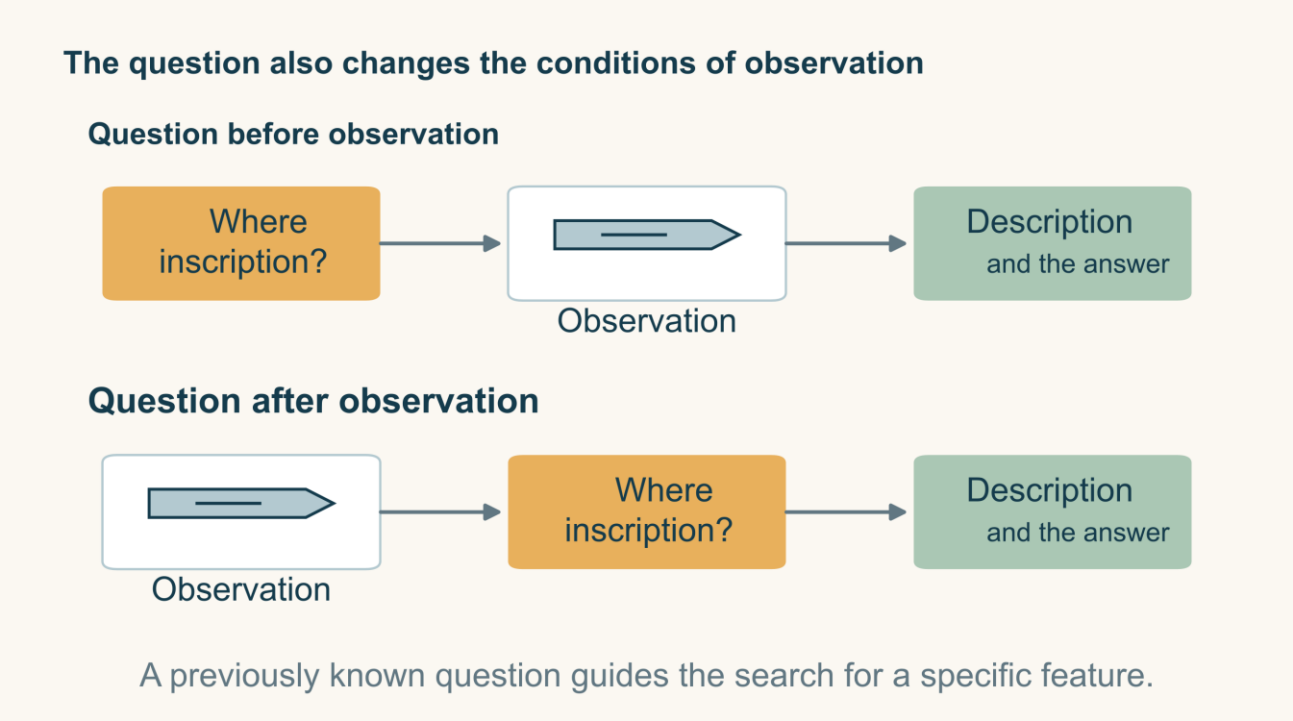

Figure 14. Question time as a condition of the procedure. The upper procedure informs the question about the inscription in advance, the lower one - after observing the object. In the second case, the moment of the question changes; therefore, the results cannot be explained only by differences in the properties of the object, ignoring the direction of observation.

Practice 6 specification of one measurement

Why do it: Create a measurement specification with explicit conditions and limits of inference.

What you will need: Distinguishing between an indicator and an explained phenomenon from chapter 5. Consider question time and answer time.

Time reference: 20–25 minutes. Can be done intermittently; This is not a speed standard.

1. Choose an item with color and a distinctive detail. Note in advance that you already know the following questions; This is an educational demonstration of the procedure.
2. Watch for about fifteen seconds, remove the object and write down the episode you experienced. Then answer separately about the color and placement of the part.
3. Check the structure of the item. Retain the original record and distinguish between the external fact and the content of the report.
4. Fill out the five lines of the specification: presentation, instructions, moment of question, message delay, evaluation rule. Add two conclusions, one allowed and one not yet allowed.

What to keep: A transcript of the episode, checkable answers, and a procedure sheet with one suggestion for improvement.

Hint: The answer “I don’t remember” is not the same as reporting no experience. Clarify which part of the chain remains unknown.

Research continuation: On new subjects, match the question before and after viewing. First, note what difference you expect and why this comparison simultaneously changes the focus of the observation.

Analysis and answers - Appendix F, chapter 6, p. 282

Self-test questions

1. Why can’t the answer “I don’t remember” be equated with the absence of experience?
2. What does the correct distinction between two objects directly demonstrate?
3. Why is the agreement between report, behavior and physiological measure sometimes weaker than it seems?

The specification shows the device of one procedure. Now we need to organize the comparison of procedures so that the possible causes of the result do not all change at the same time. This is the task of chapter 7.

Chapter 7

How the check works

In chapter 6 we described how a single indicator is obtained. Now consider two procedures and ask whether their difference can be attributed to the chosen cause. If, along with the method of work, the difficulty of the material and the order of execution change, the result allows for several explanations. The purpose of this chapter is to construct the comparison itself. We will prepare it before collecting data and keep a separate place for ideas that arise after reviewing the results.

Search for a possible refutation

A hypothesis becomes working when it is clear which possible outcome does not fit it. The statement “experience has something to do with information processing” is too broad for a single test. The statement “the stored pattern of temporal relations helps answer new questions about the text with the same total preparation time” already sets the conditions, the result and the possible comparison. You may experience improvement, no noticeable difference, or worsening.

Validation begins before data collection. The researcher determines what to change, what to measure, and what other changes might explain the result. The condition being changed is called the independent variable, and the result being measured is called the dependent variable. For example, the presence of a schema is a condition, and the proportion of correct answers is an outcome. These titles describe the roles in the study; they do not prove the independence of variables and the causality of the relationship.

Operationalization is the translation of a general concept into a specific procedure. “Text Comprehension” can be assessed with questions that require you to infer a new attitude from what you read. But the chosen procedure covers only a certain aspect of understanding. If the answers are allowed to be sought in the diagram, the solution to the problem with external support is measured. If the schema is removed, another situation is explored, involving the storage of information in memory.

Why a bad comparison is convincing

Let's present a training project. First, a person reads a complex text without notes and answers questions. Then he reads a simpler text, makes a diagram and answers better. The author concludes that the scheme improved understanding. In reality, the strategy, the difficulty of the material, and the order of execution

changed simultaneously. The second attempt could benefit from familiarity with the question format.

Such an accompanying difference is called a confounding factor if it interferes with the isolation of the cause of interest. It is not enough to list possible obstacles at the end of the report. The test design should, if possible, change their association with the main condition. If simple texts are always used with schema, the effect of schema cannot be separated from the effect of simplicity.

Not all differences need to be resolved physically. Part can be distributed between conditions according to a pre-selected rule. Others measure and take into account in the analysis. However, the calculation will not correct the lack of necessary comparisons. If there was no complex text with the circuit, the model would have to rely on additional assumptions about how the circuit works on complex materials.

Improving the project step by step

Let's start with two sets of short texts A and B. They should be similar in length, format and required output. It is advisable to prepare them before the main test and make sure during the test run that the instructions are clear. The trial run is intended to correct the procedure; its results cannot be discreetly mixed with data from testing an already modified plan.

Some participants work with the scheme on set A and without it on set B; others get the opposite combination. Within these options, the order of conditions is changed. This balancing of order and material does not make the texts absolutely identical. It distributes their features so that one set is not inextricably linked to one strategy.

Random assignment of options helps avoid choosing a condition based on the characteristics of the participant. For example, you cannot automatically give a diagram to all more experienced readers and then attribute their advantage to it. In a small group, randomness does not guarantee exact equality of composition, but it creates an understandable procedure for distribution. It is useful to combine it with a thoughtful arrangement of materials.

If the study is carried out only on yourself, you can change conditions and materials according to a predetermined schedule. The result then primarily relates to your sessions. Repeated reading of the same text creates knowledge transfer, so it is better to use new, comparable materials. A DIY project is useful for learning the procedure and finding problems; its area of generalization must remain visible.

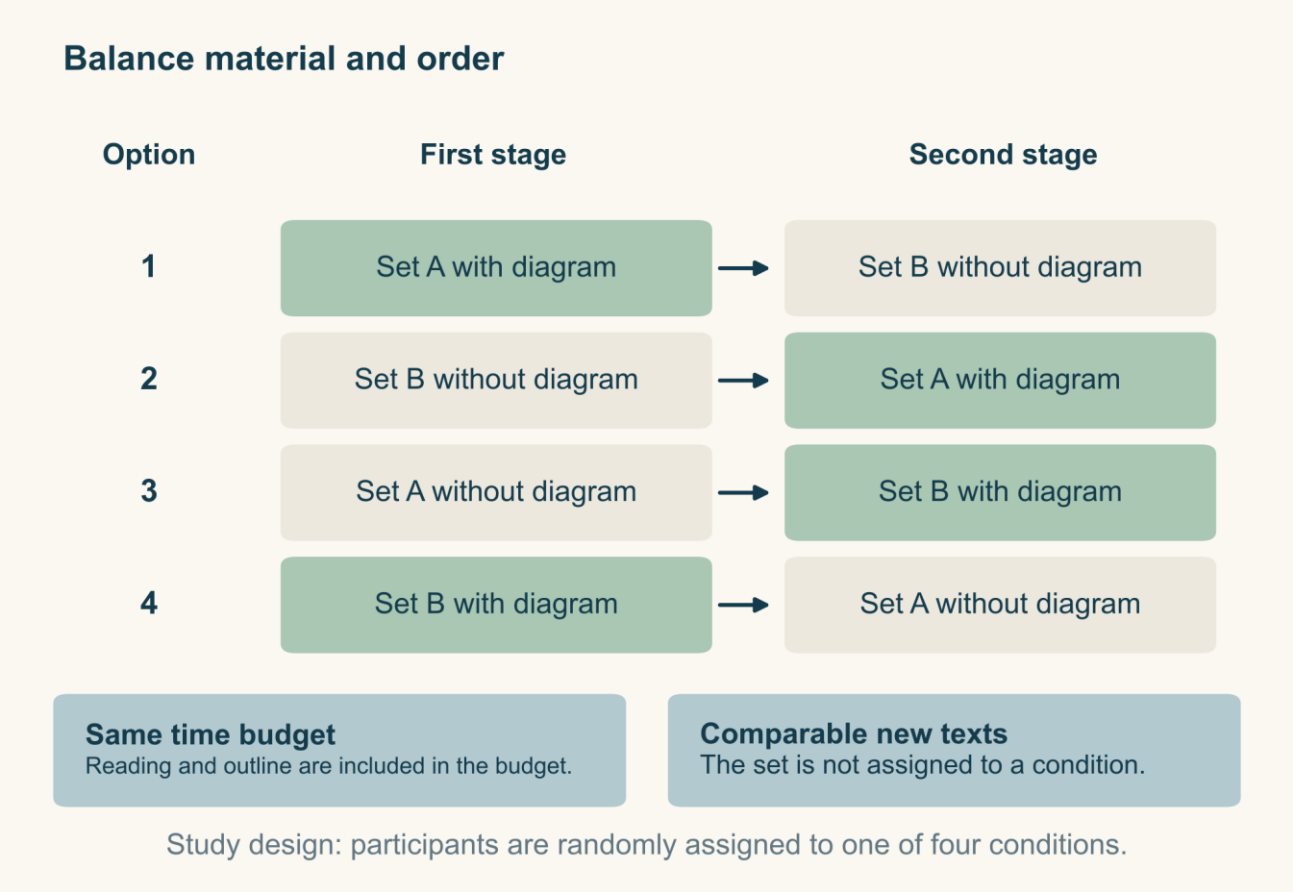

Figure 15. Distribution of material and order of conditions. The four options intersect two combinations of material with strategy and two execution orders. Each set occurs with and without a diagram, in the first and second stages. Random assignment of options and equal preparation budget complete this balancing act.

Make the result uniquely computable

Let’s assume that the main indicator is the percentage of correct answers to questions about time order. For each question you need a key in advance. If the answer is free, you should determine which formulations count. Otherwise, after viewing the results, you will be tempted to evaluate inconvenient cases more strictly and expected ones more gently. If possible, the evaluator should not see in what condition the answer was received.

We will set the total preparation time budget to be the same. This answers the practical question of whether the scheme helps when time is limited, but does not mean the same time for direct reading: part of the budget will be spent on writing. Now the negative result can be interpreted meaningfully. It is possible that the benefits of recording did not compensate for the costs of the chosen duration. The next hypothesis could concern the dependence of the effect on the time available.

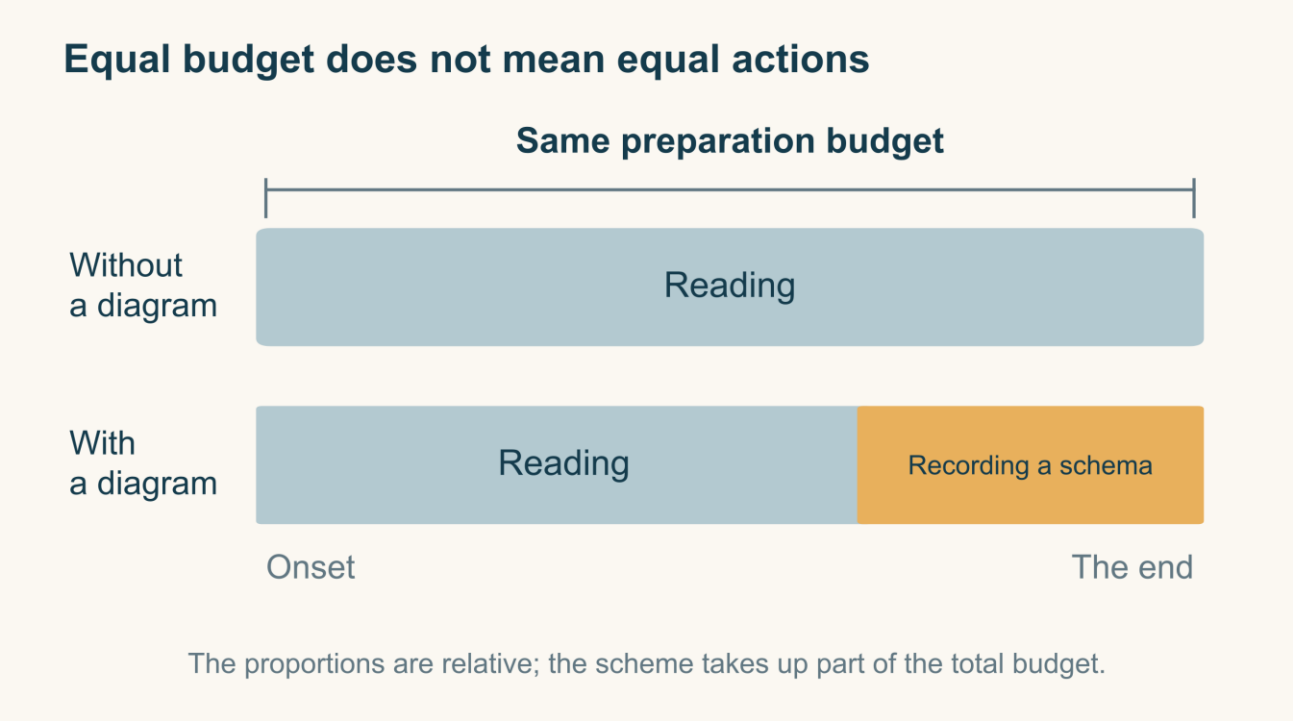

Figure 16. Overall budget and distribution of preparation time. With the same total budget, preparation with the scheme includes the time of its recording. This comparison evaluates the benefit of the procedure in a limited time, but does not equate the duration of the actual reading; The proportions of the parts in the figure are relative.

In addition to correctness, one can collect confidence and a description of the experience of clarity. But it is better to choose the main result in advance. If any of the ten indicators are allowed to be declared as evidence after review, the verification procedure loses certainty. Additional indicators remain useful for finding new questions if their research role is named honestly.

Plan before data and study after data

Brian Nosek and colleagues describe preregistration as a way to fix questions and a plan for analysis in advance. Its methodological function is to distinguish between the verification of a prediction and the explanation of the result already seen. The article does not promise that recording will correct a weak theory or poor measurement; it is devoted to the structure of the research process. [10]

For a teaching project, a dated outline is sufficient at first, which you keep unchanged next to the results. This is training in the same discipline, and not a replacement for the formal registration procedure. In the plan, indicate the question, conditions, main indicator, rules for eliminating ineligible records, number of sessions planned, and method of comparison. Record changes separately along with the reason.

Research after reviewing the data is also necessary. You may notice that the diagram only helps on texts with four events. This is an interesting new hypothesis.

The next step is to test it on new texts, and not present it as the only pre-expected prediction. Scientific work includes both the discovery of patterns and their subsequent verification.

What changes when conditions are balanced?

Let's imagine that the scheme is always used with a second, lighter text. Even an accurate measurement of success will not separate the benefits of the scheme from simplicity and training. Four distribution options break this connection: A with the scheme, then B without it; reverse order of the same pair; A without a diagram then B with a diagram; reverse order of the second pair. Now every set and every order encounters both ways of working. Random assignment of options does not guarantee the identity of people in a small group, but it does not allow the researcher to systematically assign a convenient option to expectedly strong participants.

What has now been established

The comparison plan should not eliminate all differences altogether. It must make the cause of interest discernible. The interference named after the result is useful as a constraint, but does not replace the missing control.

Practice 7 Fix a weak experiment

Why do it: Fix a project that takes increasing the length of journal entries as an improvement in the discrimination of experience.

What you will need: Measurement specification from chapter 6 and distinction between mechanism and indicator. You need two comparable sets of tasks and a pre-selected result.

Time reference: 30–40 minutes. Can be done intermittently; This is not a speed standard.

1. Formulate the result as success in distinguishing predetermined types of new episodes. Separately store the length of records as an additional indicator.
2. Set two comparable sets A and B, the same time budget and estimation rules before starting work.
3. Write down four options for the distribution of material and order, using the diagram of the worked example. Name a condition that takes into account the habit of writing or extra time.
4. For improvement, agreement, and divergent results, write down acceptable findings in advance. The project is enough for now; Gathering of participants is not included in this exercise.

What to keep: A plan with conditions, distribution, main indicator, scoring rule and three interpretation options.

Hint: The benefit of a scheme with equal total time and the benefit of a scheme with additional time are different issues. Choose one explicitly.

Research Continuation: Keep a dated version of the project. After the trial run, separate the instruction changes from the main test results and formulate a new hypothesis for the new materials.

Analysis and answers - Appendix F, chapter 7, p. 282

Self-test questions

1. Why doesn't the same preparation time mean the same preparation process?
2. What does random assignment of conditions do and what does it not guarantee in a small group?
3. How does a new hypothesis after analysis differ from a previously tested prediction?

Even a well-designed comparison produces changing numbers. In the next chapter, we will examine effect size and precision of estimation, while maintaining their distinction from causal inference.

Chapter 8

How to deal with uncertainty

The verification design already defines what is being compared to what. However, two identically designed launches can yield different success rates. We need to understand how much the observed difference limits the possible sustained effect. Here we will separately consider three questions: what happened in this series, how accurate is the assessment and what reason does the design allow us to discuss. These differences will then be needed to compare computer models and biological samples.

The difference in the table still requires evaluation

In one condition, 14 tasks from 20 were correctly completed, in the other, 16 out of 20. The shares of correct answers are equal to 70 and 80 percent. The difference is 10 percentage points. This is a description of the observed effect. It does not yet say how accurately the difference is estimated, whether it will persist in new tasks, and whether it is caused precisely by the changed condition.

Percentage points can be conveniently distinguished from relative change. The transition from 70 to 80 percent gives an increase of 10 points, and relative to the original 70 percent - approximately 14 percent. Both forms may be mathematically correct, but they answer different questions. For a teaching project, report the original quantities and proportions first so that the meaning of the change can be reconstructed.

Effect size, uncertainty, and causal interpretation are three separate issues. A large difference may not be accurately assessed. A small difference can be measured very accurately. Even the exact difference between groups does not determine its cause if the comparison device allows for systematic alternatives. chapter 7 was responsible for the quality of the comparison; here we learn to understand the variability of the outcome.

Random scatter training model

Let's imagine a generator that independently produces the correct answer in each trial with a probability of 70 percent. Let's create two conditional modes A and B with exactly the same rule. We will conduct twenty tests in each and repeat this pair six times. The following table contains synthetic data obtained for the book by a software random number generator. These are not the results of experience with people.

Repeat	Correct in A	Correct in B	Difference between B and A in percentage points
First	18	17	minus 5
Second	14	14	0
Third	13	15	plus 10
Fourth	19	15	minus 20
Fifth	11	13	plus 10
Sixth	10	14	plus 20

Here we know the generation mechanism: mode B has no advantages. However, individual comparisons create the impression of advantage or disadvantage. This makes it possible to understand random estimation error without suggesting that someone miscounted. All numbers can be written accurately, and the estimate of the stable difference will still fluctuate.

The specific spread of the table should not be transferred to any study. It depends on the number of trials, the probability of success, and the assumption of independence. In a real problem, one person's answers may change together due to learning, series complexity, or fatigue. Therefore, the generator is a tool for understanding a certain randomness model. For real data, the model has to be justified.

What Precision of Estimation Means

Estimation becomes more accurate when the procedure better constrains the range of plausible values for the effect of interest. Additional independent observations and reduction of unnecessary measurement variability are often helpful for this purpose. However, repeating the same thing is not enough if all repetitions have one common error. A thousand measurements with an incorrectly configured device does not correct its setting.

Confidence intervals are found in reports. In the frequency approach, the level 95 percent characterizes the construction procedure: when it is repeatedly applied under the conditions of the adopted model, approximately 95 percent of the resulting intervals must cover the fixed true value. It is not an automatic probability that the true value lies within a specific interval that has already been calculated. Such a probabilistic formulation requires a different, explicitly defined approach.

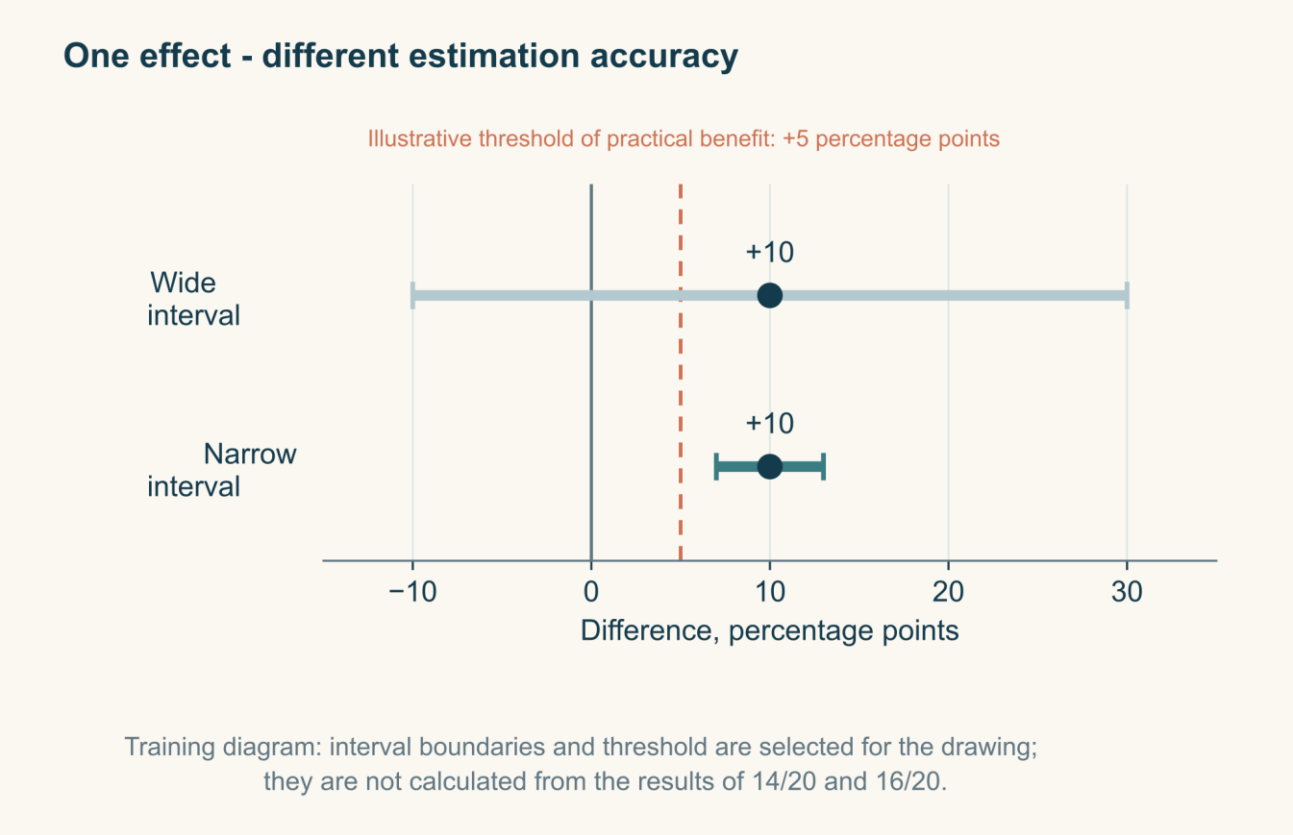

Figure 17. The magnitude of the effect and the accuracy of its assessment. In both training scenarios, the point estimate is +10 percentage points. The wide interval from −10 to +30 allows for deterioration and improvement; narrow from +7 to +13 is located above the conditional threshold +5. The boundaries and threshold are chosen for explanatory purposes and are not confidence intervals calculated from the chapter table.

At the first stage, it is enough to ask a practical question: are the data consistent with both significant benefit, lack of benefit, and deterioration under the adopted method of assessment? If so, the outcome remains uncertain for the choice between these options. If wide range is excluded, the conclusion may be meaningful even in the absence of the expected improvement. Precision allows one to constrain theories rather than just announce successes.

A practically significant quantity also needs to be defined meaningfully. If a scheme requires additional time and significantly complicates the work, the increase in one correct answer per thousand tasks may not justify its use. In another task, even such an increase will be useful. Statistical precision does not make this decision for the researcher: it shows how well the effect is known, and the assessment of benefit relates it to the purpose of application.

For hypotheses about mechanisms, the utility may also lie in small but specific differences. For example, two models solve common problems in the same way, but react differently to a certain delay. If this predicted difference is reliably measured, it helps to choose between mechanisms, even when it makes little difference to

overall success. Therefore, the criterion of content importance should be formulated together with the research question.

Don't confuse the number of answers with the number of independent cases

Let's say one person completed one hundred tasks, and another project collected one answer from one hundred people. Both files have a hundred lines, but they answer different questions. The first better describes the variability of a given person's responses in the selected series. The second contains information about differences between people, although each person is measured very briefly.

If all one hundred responses from one participant are considered as one hundred independent confirmations of the properties of people in general, uncertainty can be significantly underestimated. It is necessary to determine in advance the unit of output: test, session, person, organization of the model. For computing projects, a similar problem arises when multiple outputs from a single trained network are mistaken for multiple independently generated systems.

This doesn't mean repetitions are useless. They help assess stability within a selected unit. It is important to preserve the structure of the data: which responses relate to one participant, run, or set of materials. The analysis can then take the dependency into account instead of hiding it in a flat table.

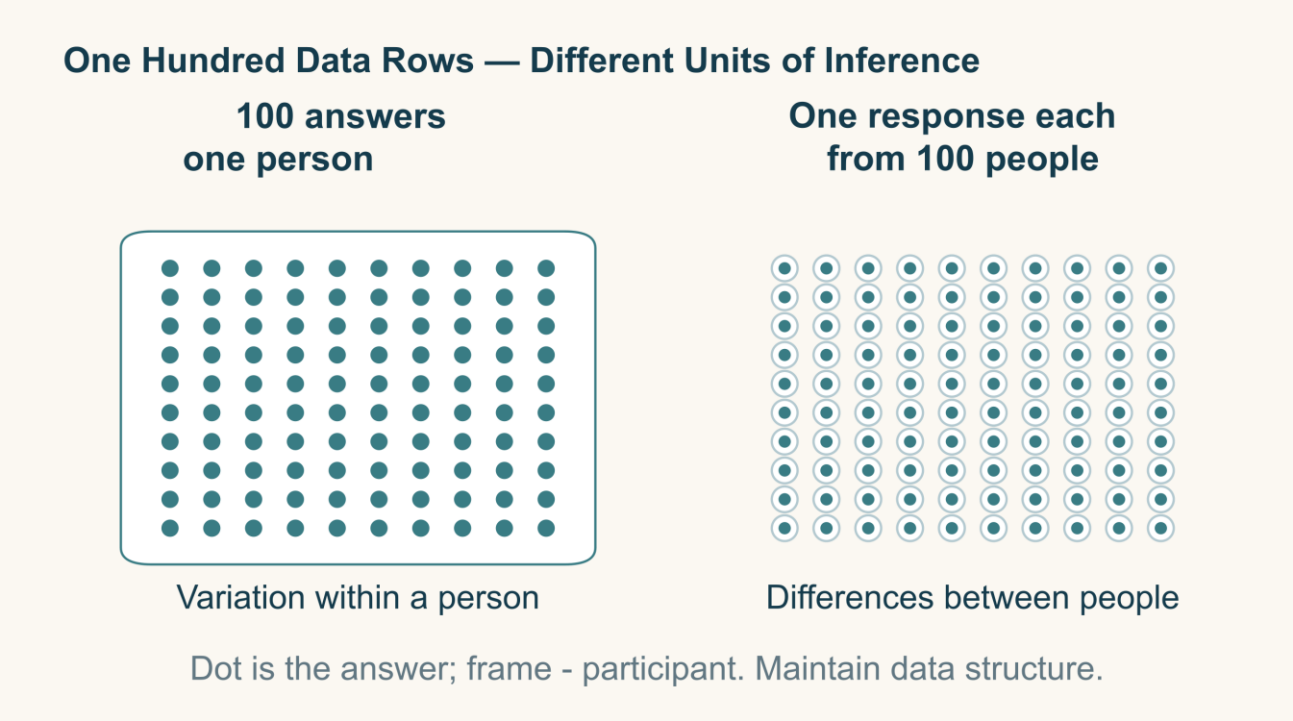

Figure 18. Responses and participants as different levels of data. On the left, one hundred answers belong to one person, on the right, each answer comes from a separate participant. The same number of rows does not make these plans interchangeable: they provide information about variability at different levels.

How Multiple Attempts Change the Meaning of a Finding

Imagine a researcher who tried twenty measures and reported only the most striking difference. Even if every calculation is done correctly, the resulting history does not show the entire search. The reader sees the result as if it was what was expected from the very beginning. The assessment of surprise must take into account the possibility of choosing a successful option from many.

Joseph Simmons, Leif Nelson, and Uri Simonsohn used simulations and experimental examples to show how the flexibility of collection and analysis can produce compelling-looking statistical results for false hypotheses. Their work links the problem to undisclosed research solutions and suggests increasing procedural transparency. [11]

The p-value, when you encounter it in a paper, is not the probability that the hypothesis is false. It measures how unusual data at least as extreme as those observed would be under the specified null model and procedure. A small value therefore does not replace an estimate of the effect, checks of assumptions, or an account of the search performed. For your first research project, a clear plan and complete results are more useful than a single impressive number.

The same score may allow different solutions

Let the improvement score be ten percentage points in two conditional scenarios. The interval from minus ten to plus thirty allows for both deterioration and significant benefit. The interval from plus seven to plus thirteen limits the possible values much more strongly. If an improvement of at least five points was considered useful in advance, the second scenario answers this question more definitely. The boundaries in this example are chosen for explanation purposes and are not calculated from the table 14 out of 20 and 16 out of 20. The visual diagram cannot be used as a result of statistical analysis of these forty responses.

What has now been established

We have learned to distinguish between the size of a result and the accuracy of knowledge about it. A large number of rows only helps in considering which rows are truly independent and what question is being asked of the evaluation.

Practice 8 Three different conclusions

Why do it: Separate the description of one series, the conclusion about the generation mechanism and the plan for the next test.

What you will need: Shares, percentages and a comparison device from chapter 7. The confidence interval is first considered as a characteristic of the estimation procedure; there is no need to derive its formula for practice.

Time reference: 20–30 minutes. Can be done intermittently; This is not a speed standard.

1. In the synthetic table of the chapter, find the fourth repetition: A 19 out of 20, B 15 out of 20; sixth: A 10 out of 20, B 14 out of 20. Calculate the differences B minus A in percentage points.

2. Write down the story that would emerge from showing just one of these lines. Then explain why it doesn't set the oscillator mode advantage.

3. For a large imprecise effect, an accurate estimate around zero, and differences between materials, suggest one of each of the following actions.

4. Specify the unit of independent repetition for one hundred responses of one person and for one response of one hundred people. Retain these differences in the analysis plan.

What to keep: Two differences, limiting their interpretation, and three reasonable next steps.

Hint: For the share 19 out of 20 the result is 95 percent. Subtracting percentage shares gives percentage points, not relative growth.

Exploratory continuation: Select in advance the quantity of practical interest and the accuracy required to distinguish it. Describe which dependent observations cannot be considered separate replications of the entire study.

Analysis and answers - Appendix F, chapter 8, p. 283

Self-test questions

1. Why is the difference between 80 and 70 percent equal to ten percentage points, and not necessarily ten percent relative growth?
2. Why isn't one hundred responses from one model the same as examining one hundred independently trained models?
3. Could the lack of expected improvement be a meaningful result?

Now the comparison has a clear indicator and limits of uncertainty. Let's move on to a mechanism that itself produces a sequence of results: the internal state will connect one moment to another.

Part 3

Mechanisms and models

Chapter 9

From nerve cell to system

Previous chapters taught how to check explanations. Now let's build a small explanation, the work of which can be traced in its entirety. The same input should sometimes produce different responses because the system stores history. The threshold accumulator will allow you to see this without knowledge of neurobiology and complex mathematics. Its connection with the nerve cell is limited to selected properties; further questioning of experience will require new relationships and arguments.

Start with a changing state

So far we have discussed how to describe experiences and test statements. Now let's move on to systems that process signals. The big new question is: How does the input at the moment, together with the internal state, determine the next step? It is suitable for both the biological process and the curriculum. The answer helps explain why the same input sometimes produces different results.

Neuron is a cell of the nervous system. Its membrane separates the internal environment from the external one, and the distribution and movement of charged particles is associated with the electrical properties of the cell. Many neurons transmit signals using action potentials—short changes in membrane voltage that propagate throughout the cell. To understand this chapter, it is enough to see a neuron as a physical system with inputs, internal dynamics, and effects on other cells.

This is not yet a model of the entire behavior of a neuron. Different cells and their parts have different properties; interactions involve chemical processes and multiple time scales. The word “neuron” in the name of a software element usually denotes a specific abstraction. Therefore, when reading a model, ask which properties are retained and which are omitted for the sake of the chosen task.

How quantitative explanation differs from imagery

Alan Hodgkin and Andrew Huxley constructed a quantitative description of nerve fiber membrane currents by relating changes in conductance to electrical

excitation. Their work shows a research transition from measurements to a system of rules capable of reproducing the dynamics of a process. This is a model of certain electrical properties of a nerve cell; it was not offered as evidence for the emergence of subjective experience. [12]

A useful image like “neuron collects neighbors' votes” helps get started, but hides important questions. How long does the influence of input last? Does the next answer depend on the previous one? Is there a threshold? What happens after the answer? A computable model forces these questions to be answered with sufficient certainty that the result can be obtained without a new solution from the author at each step.

Our first model will be much simpler than the biological one. It is needed to study accumulation, leakage and threshold. Its numbers are arbitrary and are not measured in millivolts. Similarity with certain properties of a neuron does not give the right to transfer all the properties of a living cell to it.

A worked example of a threshold accumulator

Let's imagine a state variable initially equal to zero. At each step, 60 percent of its previous value is retained and a new input is added. If the resulting number reaches one, the system issues a pulse and resets the state to zero. If the threshold is not reached, there is no pulse, and the state persists until the next step with subsequent leakage.

Let's provide inputs 0.6; 0.6; 0.6; 0; 0.6. At the first step there is no previous accumulation, so the new value is 0.6. On the second, 0.36 remains from the previous state and 0.6 is added: it turns out 0.96. On the third, 0.576 is saved and another 0.6 is added. The value 1.176 crosses the threshold, a pulse occurs, after which the state is zero.

In the fourth step, the input is zero, so the state remains zero. On the fifth it again reaches 0.6. The same inputs of the first, second and third steps led to different states because the history was different. There is no additional “when to respond” decision in the program: the result is obtained from the update rule.

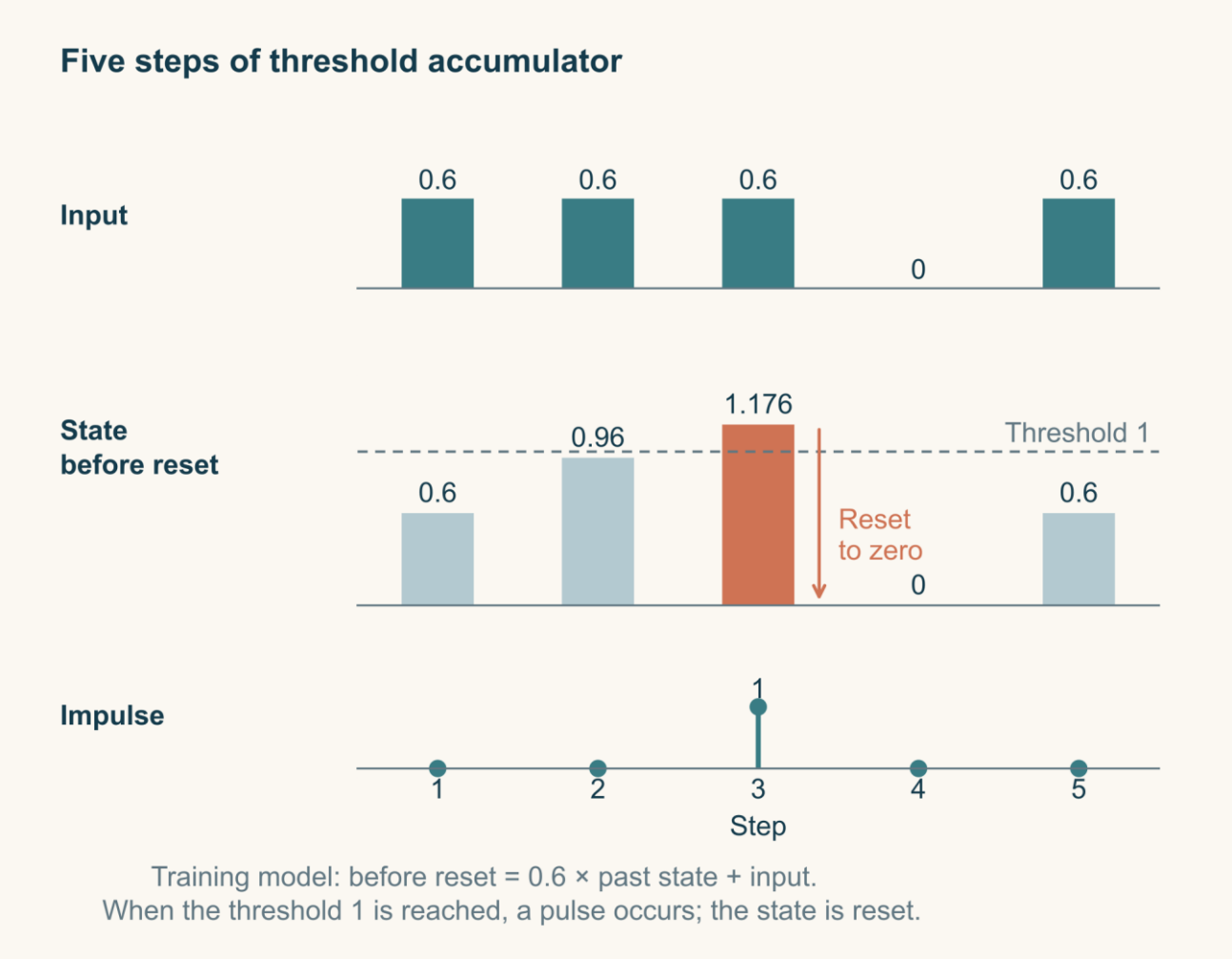

Figure 19. Dynamics of accumulation and discharge. Exact results of the program from the chapter. Inputs 0.6; 0.6; 0.6; 0; 0.6 give states before resetting 0.6; 0.96; 1.176; 0; 0.6. Only at the third step the threshold 1 was reached. All values are arbitrary and not expressed in millivolts.

Here is the complete computable version in Python. The word state denotes the state, value is the next input, spike is the presence of a pulse. The period is used in the code as a decimal separator. The indentations indicate which actions are repeated for each input and which are performed only when a threshold is crossed.

```
inputs = [0.6, 0.6, 0.6, 0.0, 0.6]
state = 0.0
for value in inputs:
    before_reset = 0.6 * state + value
    spike = before_reset >= 1.0
    state = 0.0 if spike else before_reset
    print(round(before_reset, 3), int(spike), round(state, 3))
```

The program displays the value before a possible reset, then 1 if there is a pulse or 0 if there is no pulse, then the final state. It can be done or checked with a pencil. In both cases the rule must produce one sequence. This transparency is useful

before moving to large models whose internal workings can no longer be traced in their entirety by hand.

From one element to a network

Now let's imagine several accumulators. The output of one is included in the update rule of the other. The coupling coefficient determines the strength of this influence. A positive contribution can bring the state closer to the threshold, while a negative contribution can move it further away. This is a textbook diagram of excitatory and inhibitory influences, not a universal description of all biological interactions.

If connections go only from input elements to output elements, a directed processing sequence is obtained. If a signal can return through several connections to a previously participating element, feedback occurs. It allows previous results to influence further processing. But feedback does not necessarily support stable activity: depending on the rules, it can attenuate, intensify, fluctuate, or lead to other modes.

It is important to distinguish between a change in activity and a change in the system itself. Our state changes at every step, but the coefficient 0.6 remains constant. If you add training, the coefficients can also change according to a separate rule. Then the model will have two types of history: a short-term history of the current state and a longer change in parameters. Mixing these levels makes it difficult to understand what exactly the model remembered.

A network does not become an explanation just because it contains many elements. It is necessary to show what problem their organization solves. For example, one network can store a signal between events, another can select a sequence of a certain type, and a third can select an action. These differences can be tested before asking questions about the organization's relationship to the experience.

A number or impulse receives research meaning through its relationship to conditions and use. If an element responds more often to red images, you can test whether the relationship remains the same across changes in shape, brightness, and environment. Then you need to find out if another component uses this response to distinguish color. The name "red detector" would be premature if the element actually responds to a luminance that happens to be the same as the color in the training set. The presentation of content requires verification, not just a well-chosen label for a network node.

One step consists of several actions

If the previous state is 0.96, first save 0.6 from it: we get 0.576. Then we add a new input 0.6 and get 1.176. Only now we check the threshold 1. It has been reached, so we record the impulse and reset the state to zero. The value before reset and the state after reset are different. If you change the order of actions, you get a different model. This is why the resulting line does not replace an explicit transition rule.

What has now been established

The accumulator shows the causal contribution of the stored state. The impulse arises from the rule, not from an additional internal decision. Now let's check what value the distribution of the same sum of inputs has over time.

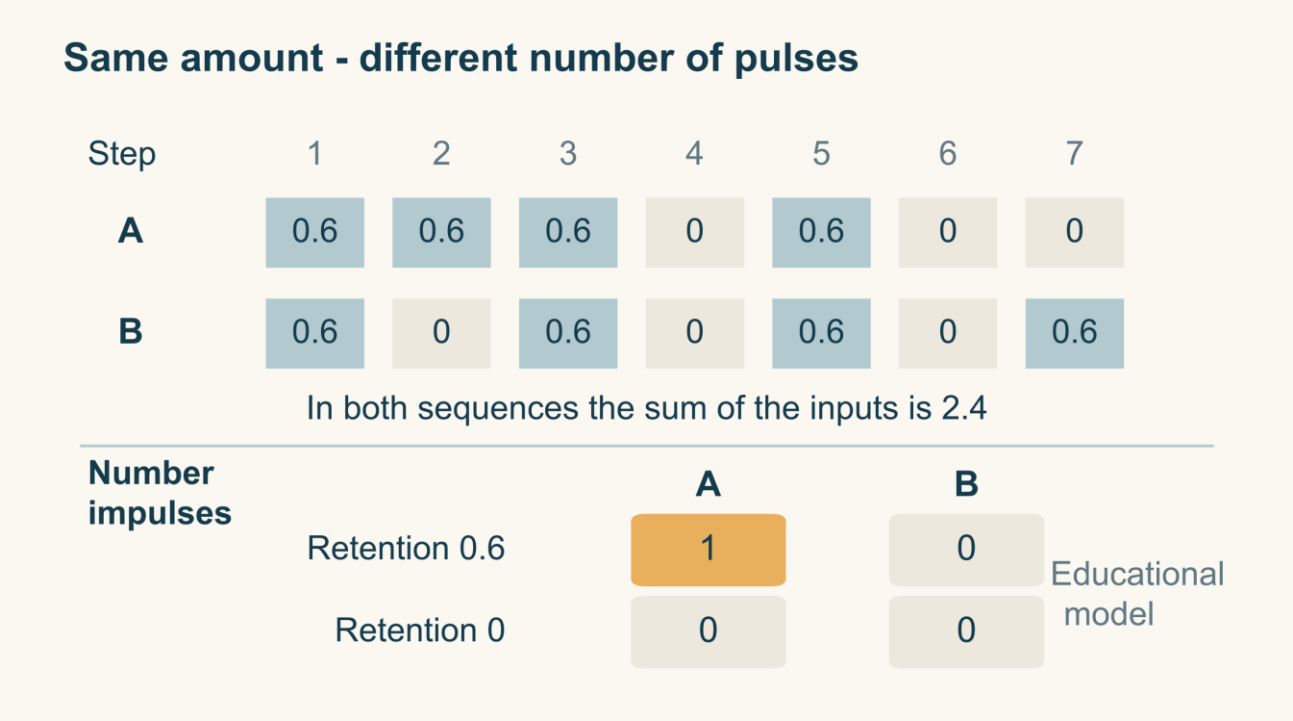

Figure 20. Distribute inputs and disable state saving. The two practice sequences have the same sum 2.4, but when 0.6 is stored, only the first one causes one pulse. At zero conservation, there are no impulses in both series: the intervention highlights the role of state history in this training model.

Practice 9 Same amount different times

Why do it: Check the explanation through temporary accumulation in comparison with the explanation through one total amount.

What you will need: Ability to multiply decimal numbers. The input comes from outside, the state is saved between steps, the output is obtained after updating.

Time reference: 20–30 minutes. Can be done intermittently; This is not a speed standard.

1. For A use 0.6; 0.6; 0.6; 0; 0.6; 0; 0. For B use 0.6; 0; 0.6; 0; 0.6; 0; 0.6. Initial state zero, saving 0.6, threshold 1.
2. Before calculating, predict the number of pulses. For each step, write down the input, the value before reset, the pulse and the state after reset.
3. Execute the same rule for both sequences manually or with the code from the chapter.
4. Repeat with zero remaining and explain which difference disappeared and why.

What to save: Two step-by-step tables, a preliminary forecast and a conclusion about the role of gaps between inputs.

Hint: After the impulse, the next step starts from scratch. Do not transfer the value 1.176 as a saved state.

Exploratory continuation: Propose a new pair of equal-sum sequences on which the sum and time structure versions yield different expectations. Fix the rule before calculation.

Analysis and answers - Appendix F, chapter 9, p. 283

What step has been taken towards researching experience?

We received a minimal language of dynamic models: input, state, transition rule, output and parameter. It can be used to ask what the system distinguishes, what it holds, and how it uses the past. Later we will add views of sources, assessment of reliability and own actions. Each component will be introduced for the sake of explainable behavior and testable prediction.

The presence of impulses or feedback does not in itself resolve the issue of qualia. However, it is now possible to formulate a hypothesis more precisely than “complexity gives rise to consciousness”: name a certain organization, indicate the expected relationships between its states and the properties of experience, then look for conditions for distinguishing alternatives. This is the preparatory role of simple models.

Self-test questions

1. Why can the same input in the first and third steps produce different outputs?
2. How is changing state different from learning parameters?
3. What does the comparison of sequences with the same sum and different intervals between inputs check?

The state has learned to hold on to the past. The next question is what exactly it represents and how the context changes the use of the same input.

Chapter 10

Signal processing and access to their content

The drive showed how history changes the answer. This is not enough for perception: it is necessary to distinguish between the content, the method of interpretation and access to the result. The same oval can be used as a letter or number depending on its neighbors. In this case, we will analyze direct and repeated processing. When comparing them, we will make sure that the architecture does not gain an advantage only due to additional inputs or time.

One image and several research questions

You look at the page and look for the section number you need. It simultaneously contains letters, margins, punctuation marks and a small signature. Having all these elements in front of your eyes does not mean that each of them has the same impact on the current task or is equally available for subsequent description. To explore the situation, you need to separate several concepts.

Information processing means changing system states depending on input and internal rules. Attention in an operational functional sense refers to the selective allocation of processing: certain sources or cues are favored over others. Access means the ability to use content in specific actions, reasoning, or responses. A report is a specific action of a message. Experience refers to how a situation is subjectively given.

These concepts may relate to the same episode, but are measured differently. We can check if the answer to the section number has changed after the hint. We can ask what was clearly visible. We can find out if the location of the signature has been saved for the next task. If all the results are combined with the word “awareness,” it becomes difficult to understand what process the model explains.

Why does the length of the report depend on the procedure?

George Sperling examined brief presentations of character arrays, comparing the full report with the report of the portion specified after presentation. The partial report advantage and its dependence on pointing delay indicated that the availability of information after a brief presentation cannot be assessed solely by the number of sequentially named symbols. The experiment raises the question of storage and communication limits; by itself it does not conclude the debate about the full scope of phenomenal experience. [13]

The general lesson can be understood by a common example. You notice several objects on the table, then begin to describe the first one in detail. During the story, information about others may become less accessible. The short final list then depends not only on the initial processing, but also on timing, message order, and detail requirements. In a specific study, the contributions of these processes should be separated.

The opposite situation is also possible: the correct answer may be based on information that a person finds difficult to describe in words. Therefore, the question “how much was told” is useful for studying the report, but requires additional justification as a measure of all available content. It is better to clarify for which operation the information was available and at what moment.

Attention changes the task

Daniel Simons and Christopher Chabris examined noticing an unexpected event while observing a dynamic scene in an attention task. The probability of its detection depended on the design of the main task and the properties of the unexpected object. The work demonstrates the limitations of the observation under a particular load; the absence of a subsequent message should not be made into a universal criterion for the absence of any processing. [14]

Attention is conveniently understood by asking what information the system prefers to use now. When searching for a section number, the shape of the numbers may be more useful than the content of the adjacent paragraph. When checking typos, the priority changes. The same external elements are involved in different computational tasks depending on the purpose.

However, attention and consciousness do not become synonymous just because they are closely related. First you need to select an attention indicator, an experience indicator and the conditions for comparing them. If attention is defined through search success, and experience through the clarity report, the relationship between these indicators is examined. A substantive theory should explain their similarities and possible discrepancies, and not eliminate the difference by renaming.

An example of an ambiguous sign

Imagine a drawn oval, which can represent the letter O or the number 0. The drawing itself is the same in the two training cards. On the first there are K and T around him, on the second - 1 and 4. In the first environment, the letter version is natural, in the second - the numbers. Context adds information about what role the sign is likely to play in the record.

Let's build a simple model. The first stage evaluates only the central sign and stores two options. The second one considers the neighbors. If both neighbors are letters, the letter version takes precedence; if both digits are digital. The third displays the selected category. Here you can indicate what information was received at each stage and why the result differs with one central figure.

This is the author's teaching example, and not a reproduction of a specific experiment. It shows how context can be involved in the interpretation of input. Whether the subjective vision of the form has changed or only its designation has changed requires a separate question to the observer. The same response “letter” can be accompanied by different descriptions of experience, and the category model does not yet explain these differences.

One pass and re-refinement

In direct processing, information goes through a sequence of transformations to an answer. With recurrent processing, results can come back and change the operation of previously participating components. The word "recurrent" here means the presence of such a return. For example, a previously recognized context may influence further evaluation of the local shape.

In our problem, we can build a recurrent version: first store the ambiguous state of the central sign, then obtain information about the neighbors and update the previous state. But you can also create a model that immediately receives the entire card and selects a category with one rule. In the two examples described, their answers will be the same. Therefore, coincidence with people's answers does not yet determine the internal organization.

For a meaningful comparison of architectures, the amount of available information and computing resources must be taken into account. If one model was shown the entire card, and the other only an oval, the advantage of the first is explained by the difference in inputs. If the reprocessing was given more time, it is necessary to find out whether it was feedback or additional calculations that provided the gain. A good problem allows these explanations to diverge.

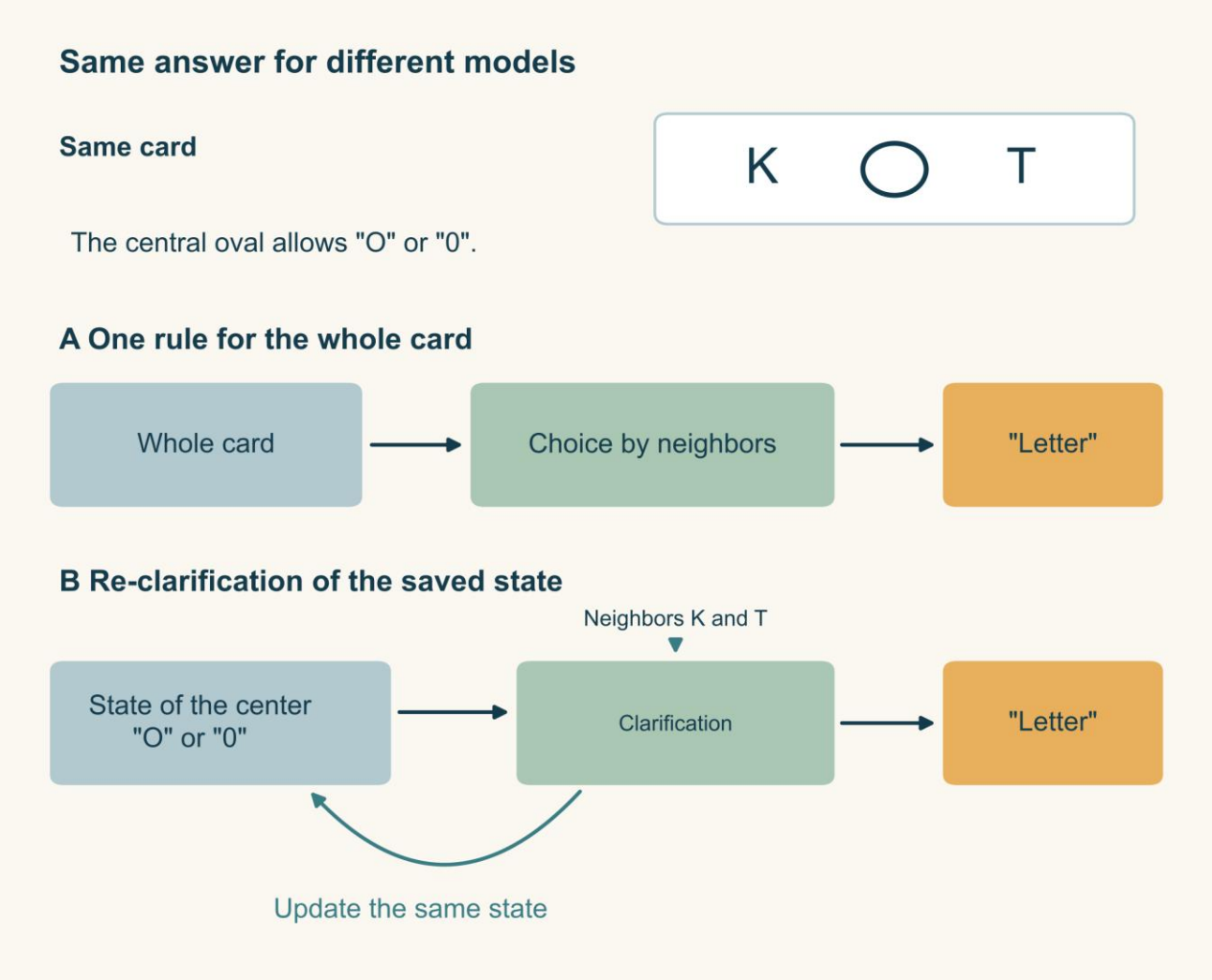

Figure 21. Two ways to use context. Both training models call the central oval a letter with neighbors K and T. The first uses the entire card with one rule. The second preserves the ambiguous state of the center, receives context and refines the previous state; the return arrow indicates this update. The equality of answers in the example does not determine the internal organization and does not prove the presence of experience.

From internal state to use

Let's say the selected category is saved within the program. One output might print the word “letter,” another might select a dictionary to continue writing to, and a third might decide whether additional checking is needed. These capabilities show different forms of access to the same processing result. The presence of an internal number does not mean that it is actually used by all components.

To test the role of a state, it can be changed or temporarily made unavailable to a specific output. If the response seal was broken, but the continuation choice was preserved, a functional discrepancy arose. In research, it is important to describe it literally. You cannot declare a system completely devoid of information just because one communication channel is broken.

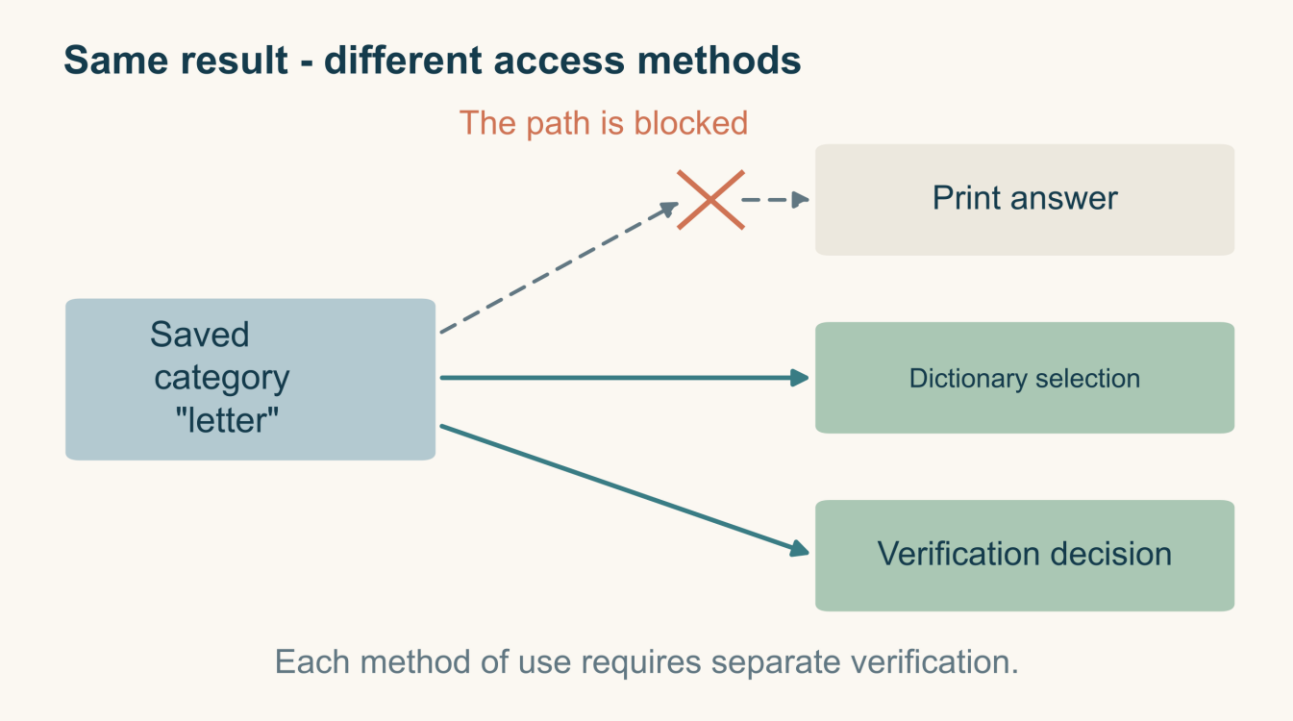

Figure 22. Selective channel disruption. In the learning architecture, the selected category is available for printing, vocabulary selection, and review decisions. Only the print path is blocked; the integrity of other uses is checked separately and is not inferred from a single missing message.

Such an analysis prepares the design of candidates for the subject. We will be able to ask what information is combined, what actions depend on it, and how the system evaluates its own decision. The relationship of this organization to experience will then become a separate testable part of the theory.

Why a hidden correct mark is not always recoverable

The same oval and digital neighbors are drawn on two cards. The author in one case conceived a number, in another a letter. If the device only sees the image, observations do not distinguish between these designs. An error regarding a hidden label here does not necessarily mean poor processing. To explore the usefulness of context, one must make an early choice between evaluating the most reasonable interpretation from the available evidence or guessing an inaccessible intention. This distinction will be useful when later auditing agent logins.

What has now been established

The correct name of the sign, the experienced form and the availability of the content for different actions are different results. Now let's compare the explanations on one explicitly given set of cards.

Practice 10 Compare three explanations

Why do it: Separate the role of context from the advantage of information and from the inaccessible environmental label.

What you will need: Entry, status and exit from chapter 9. Distinguish between an observed design and a hidden artist's mark that the device may not know.

Time reference: 25–35 minutes. Can be done intermittently; This is not a speed standard.

1. Make six cards with the same central oval: two with neighbors K and T, two with 1 and 4, two more repeat these images, but have opposite hidden author’s marks.
2. For each, write down the visible information and a separate hidden label. Set a basic rule: letter neighbors lead to the letters category, digital neighbors lead to the numbers category.
3. Compare a device that sees the entire card at once and a device that first receives the center, then its neighbors. Point out the moment when they have the same information.
4. Explain what can be determined from matching responses and what additional conditions are needed to distinguish the mechanism of re-specification.

What to keep: Six cards, two device rules, and an explicit list of available time information.

Hint: A different order of inputs is already a difference in conditions. To validate the architecture, you need to provide a competitor with comparable memory and compute.

Research continuation: Let the neighbors disappear before the center appears. Set the memory of each version and get a prediction before trying. Compare with the direct model, which is allowed the same external memory.

Analysis and answers - Appendix F, chapter 10, p. 284

Self-test questions

1. Why is the number of parts named not an automatic measure of the total initial experience?
2. Why doesn't the success of the context model prove the need for recurrent processing?
3. What needs to be clarified in the statement “information is available to the system”?

We have separated observation from hidden agenda. Now let's apply this discipline to event sourcing: how a system can evaluate whether an observed change is due to its own command.

Chapter 11

Body and action in self-reference

Context helps to interpret the sign, but in action it is also necessary to evaluate the source of the change. Did the movement coincide with the command by chance or did the command actually control it? There will be two related tasks in this chapter: integrating sensory causes and distinguishing self-effects. Before the numerical example, we will analyze probabilistic updating as a rule of version change, maintaining its difference from the experienced bodily belonging.

When several signals refer to one event

You raise your hand and at the same time see its movement, feel the change in position and get new sensations from the clothes. To describe this episode, it is not enough to say that many signals were received. It is necessary to explain why they are associated with one part of the body and one action. A similar movement on a nearby screen may remain an external event, although it also occurs before the eyes.

One approach views the problem as inferring sources. The system must evaluate whether visual, bodily, and motor information are related to a common cause or arose independently. Such an assessment may take into account temporal coherence, spatial coherence, and past reliability of sources. The word "output" here does not require an internal phrase or conscious reasoning: it denotes the operation of a model that distinguishes between possible explanations for the input.

Majed Samad, Albert Jin Chung, and Ladan Shams applied the Bayesian causal inference model to the perception of body ownership. They compared the model's predictions with experiments on the illusion of an artificial hand. The work shows how the assumption of the unification of sensory causes can be made computable and testable with data; it does not establish that any combination of signals is already an experience of one's own body. [15]



Figure 23. Hand tool and observed result. A detailed illustration of an action through an instrument: the movement of the hand is accompanied by a change in the pattern on the screen. To study authorship and self-attribution, it is necessary to clarify what information the system associates with one cause and how it verifies this connection.

A Common Cause Does Not Yet Establish Body Ownership

The sound of the impact and the movement of a falling book may also have a common cause. Therefore, the source pooling model solves a more general problem than bodily identity. To move to the body, you need to clarify which signals relate to it, how its position is represented, and what relationships connect events with the actions of the system.

Let's return to the distinction of chapter 3. Belonging answers the question whether an object is given as part of its own body. Authorship refers to the experienced or appreciated connection of an action with its outcome. The system can control an external tool and maintain a distinction between the tool and the body. It can receive information about the position of a stationary limb without performing any voluntary movement at this moment.

The work of Kalckert and Ehrsson, discussed earlier, provides experimental grounds for studying these components separately. [4] Our model will therefore have separate variables for combining sensory sources and assessing the causal contribution of the command. Even if in ordinary episodes they often change together, their separation allows the formulation of new testable conditions.

Why does evidence change the weight of a version?

Let's first distinguish between two questions. How likely is version C before the new observation? How probable is the observation itself if C is true? The first quantity is the preliminary probability of the version. The second is called the likelihood of observation under this version. They are related by calculation, but are not interchangeable.

Let sixty out of one hundred conditional cases have a right goal. The gauge is correct eighty percent of the time, regardless of the side. Then, among sixty right targets, forty-eight signals from the right are expected, and among forty left targets, eight such signals are expected. If we received a signal from the right, we consider these fifty-six cases. In forty-eight of them the target is actually to the right. The new share is equal to 48/56. The numbers here describe the selected training probability model, not a survey of people.

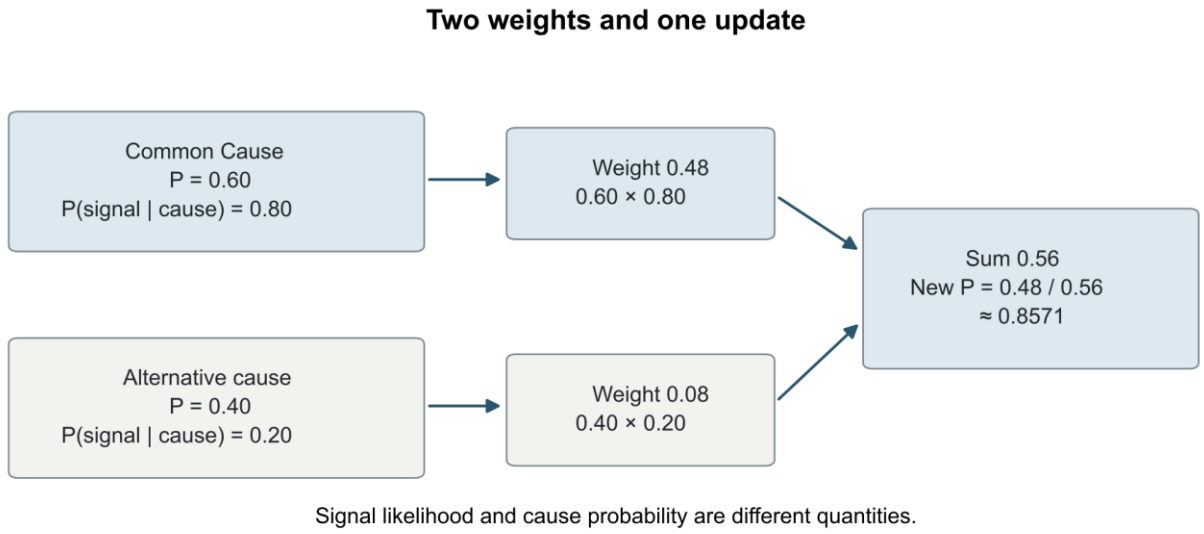

Figure 24. Consistent updating of causal versions. Training calculation with preliminary probability 0.60 and symmetric reliability 0.80. First, the unnormalized weights 0.48 and 0.08 are constructed, then normalization is performed. Numbers provide a model, not a measurement, of experience.

The same calculation can be written down for two arbitrary versions. First, we multiply the preliminary probability of each by the likelihood of the data obtained. We get two unnormalized weights. Then divide the required weight by the sum of both. Division returns the probabilities to a total of one. If the sum is zero, the data is not possible under both versions; this requires model revision or input testing rather than an arbitrary response.

The odds ratio allows you to perform the same operation compactly. The odds of C versus P are P(C)/P(P). The new feature multiplies them by the likelihood ratio P(sign|C)/P(sign|P). With initial odds of one to one, a feature that is three times more likely at C gives three to one, that is, probability C 3/4.

The following testimony should not automatically be considered a new independent voice. If one gauge simply copied another, identical numbers do not produce double the base. For sequential multiplication, the likelihood of a new feature is used, taking into account already known ones; conditional independence is a simplification that does not change this likelihood. Now the numerical example can be read as a specific mechanism and at the same time see its premises.

Worked example of two versions

Let's build our own educational model. Version C assumes that visual and bodily signals refer to the same object. Version P assumes different sources. Until a new observation, we will give the versions equal weights. Then we get a sign: the changes occurred almost simultaneously. For the exercise, let's say that such a coincidence is three times more likely with version C than with version P.

After taking into account the sign, the ratio of the weights will become three to one. Of the total sum of four shares, three fall on C: its model probability is equal to 75 percent. This is not a measured probability of hand ownership, nor is it an estimate of a real person. The numbers are chosen to show an update of two competing explanations.

Now comes the second sign: the positions of the objects diverge greatly. Let us agree that it is four times less probable at C than at P. The previous ratio of three to one is multiplied by one quarter and becomes 0.75 to one. Probability C is now approximately 43 percent. New information has changed the model's preference, although the first feature continues to be taken into account.

The calculation can be written briefly. odds means the ratio of probability C to probability P. ratio indicates how many times the next attribute is more likely at C. The last line converts the ratio into a fraction of the total amount.

```
odds = 1.0
for ratio in [3.0, 0.25]:
    odds = odds * ratio
    probability_common = odds / (1.0 + odds)
    print(round(probability_common, 3))
```

We received 0.75 and approximately 0.429. In this simplified model, the characteristics are considered independent when the version of the cause is fixed. If temporal and spatial coincidence are interrelated, they cannot be multiplied as independent evidence without verification: some of the information may be counted twice. A more accurate model should specify a joint dependence or evaluate the second characteristic taking into account the first.

What One's Own Action Adds

The passive match of signals and the result of the command create different verification possibilities. Let's say the program sends a left command and expects the pointer to move to the left. It compares the expectation to the actual change. But a coincidental result alone is not enough to establish causal control: the external recording could have accidentally performed the same movement.

Now the program changes the commands according to the new sequence. If the pointer continues to match them, the control version gains more support. If the old recording is played back regardless of the new commands, a discrepancy appears. Intervention in the sequence of actions helps to distinguish between causes that, when passively observed, might look the same.

An action model must store at a minimum a command, an expected consequence, and an observed result. It is also important to indicate the time: a late consequence may relate to a previous command. If you compare it only with the current one, the system will mistakenly attribute failure or external interference to itself. History is needed to establish relationships between events, and not simply to increase memory capacity.

At the same time, the functional assessment “this command caused the event” remains separate from the experienced sense of authorship. For human research, model estimates can be compared with reports. For an artificial system, the source discrimination itself and its use in behavior are first tested.

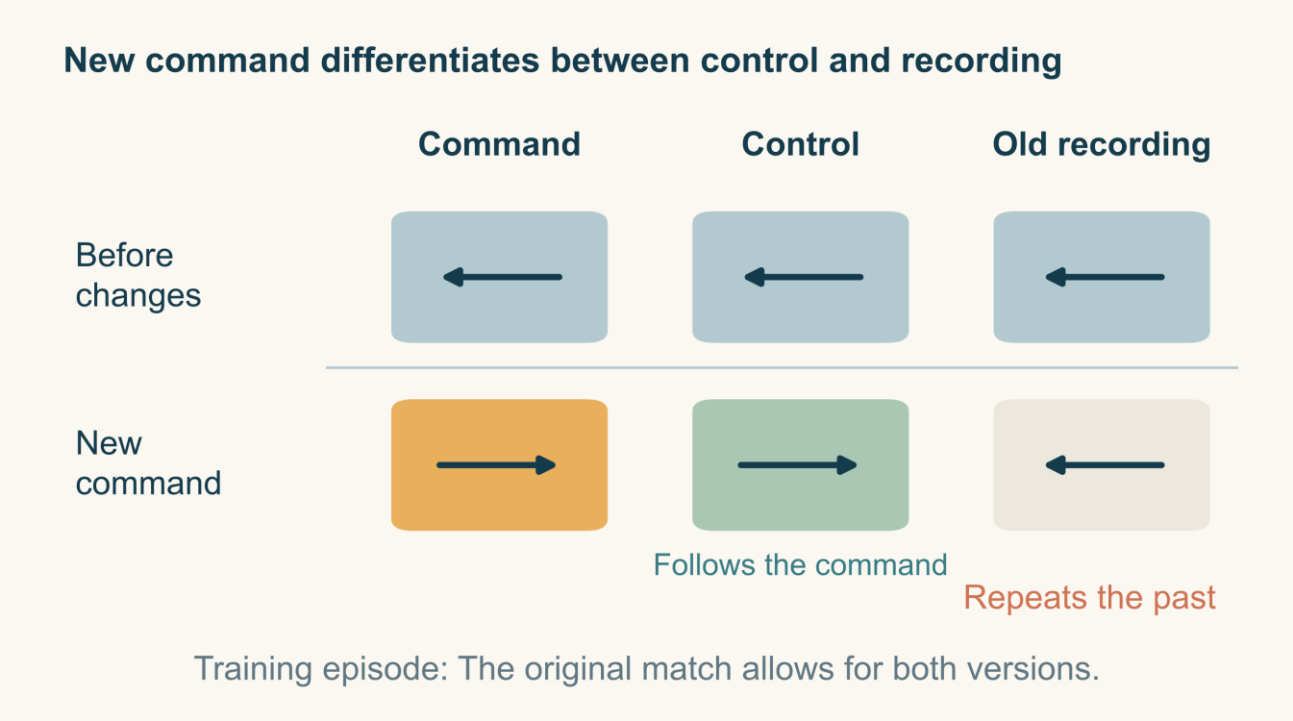

Figure 25. Distinguishing causes by changing the command. With the initial command to the left, the movement of the pointer can coincide with both the control and the external record. In the fictional sequel, a new command to the right makes different predictions: the controlled pointer changes direction, but the entry repeats the previous movement to the left.

Compare models of memory and self-reference

Let's imagine three versions of the program. The first looks only at the current match of signals. The second additionally takes into account the overall reliability history, but does not distinguish between sources. The third stores history separately for its own commands, external events and each observation channel. These models answer different questions about what information is needed for robust causal discrimination.

The third program does not gain an advantage automatically from the word "own." We must test exactly what information this label contains and how it affects prediction. If the program was simply given the correct answers, its success explains nothing. If it stores more data, it should be compared with a general memory model of comparable capacity. Otherwise, the contribution of organization becomes confounded with an advantage in resources.

A meaningful hypothesis might sound like this: separate history of sources helps to transfer the assessment of authorship to new combinations of delay and interference. The check should use combinations that were not encountered during setup. The result then shows the ability to apply the intended relation, and not just remember familiar sequences.

How a delay changes the meaning of a coincidence

At moment 1 a command was sent to the right, at moment 2 to the left. The visible result to the right occurs only at moment 3. Comparing only the current command and the current effect may mistakenly consider the result to be external. The story allows for a connection to an earlier command. But if an external recording reproduces the same series, passive observation is not enough. A new pre-selected sequence of commands allows you to check whether the result follows the current control or the old record. This control distinguishes between causal accounts and does not itself measure the sense of authorship.

What has now been established

The general cause of signals, causal control and bodily belonging did not become one indicator. We now have ways to test the first two relationships and can separately ask how their estimates relate to human reports.

Practice 11 Paper Source Researcher

Why do it: Construct a transparent analysis of the source of the event without revealing the correct reasons to the model.

What you will need: Concepts of observation, latent state, and internal evaluation. The probabilistic transition in this chapter explains the necessary calculations; There is no need to derive Bayes' rule yourself before reading the chapter.

Time reference: 30–40 minutes. Can be done intermittently; This is not a speed standard.

1. Set commands in steps: right, left, left, right. For the control version, the effect appears in two steps. For external recording, take a cyclic sequence of right, right, left, left, repeating in the same order regardless of the commands.
2. Make two observation tables and keep the true reason separately. For the first available events, write down the delay versions zero and two steps.
3. Continue the commands with a new row to the left, to the right. Before discovering new effects, write down what the control predicts and what the external recording predicts.
4. Highlight cases where there is not enough data yet, and explain which new command will actually distinguish the remaining explanations.

What to keep: Tables of commands and effects, two causal versions and a preliminary forecast of a discriminating continuation.

Hint: The first two steps of the delayed channel may not yet correspond to a command within the observed window. Don't fill them with a fictitious connection.

Exploratory continuation: Add noise and uncertain source recognition. Compare shared memory and separate histories with the same inputs and comparable capacity.

Analysis and answers - Appendix F, chapter 11, p. 284

Self-test questions

1. Why is the inference of common cause not the same as bodily belonging?
2. What makes it possible to distinguish between one's own control and a random coincidence with an external recording?
3. Why doesn't more stored data prove the benefit of self-reference?

History has been helpful in distinguishing causes. The next chapter will add another question: how reliable is your own answer and how to evaluate this before getting the result.

Chapter 12

Memory and time in assessing thinking

The system can already store information about the world and action. Now let's give its the task of predicting its own success. To do this, you need to be especially careful in maintaining the order: the forecast is written down before the outcome. A simple counter model will show how to learn from the past, and a constant probability comparison will test whether learning is useful in a particular series. Later, this analysis will become the basis for monitoring and reviewing the strategy.

The present moment depends on history

When you finish reading a sentence, its first words are no longer the focus of current attention, but may participate in understanding the end. To explain this behavior, the model must preserve information between events. In chapter 9, one changing number was enough for this. In chapter 11 it was necessary to distinguish between the history of commands and external sources. Now we will add an assessment of how well the system copes with its task.

It is useful to separate the three scales. The current state determines the nearest answer. The saved history allows you to compare the current episode with previous ones. Parameters define stable processing features and can change during training. These scales are not necessarily realized by three separate boxes. The distinction is needed to understand what change explains behavior and how long its influence lasts.

The presence of memory should not automatically be called continuity of experience. The file can store information for a long time without performing any actions. For functional research, it is important to examine whether stored content is used, how it changes decisions, and what happens when it is violated. The relationship of these processes to experienced time will be a matter for further theory.

Forecast and assessment of the answer already given

Before the assignment, you can expect to do well because you are familiar with the topic. After reading a specific condition, confidence changes. After issuing the answer, it becomes possible to evaluate the solution again, and after checking, find out its correctness. These four points contain different information. They cannot be mixed together in the same “confidence” column without time stamps.

It is especially important for the model to exclude the future cue. If the correctness of the answer is already known, it is easy to evaluate it. Of research interest is the forecast before feedback or the assessment of the answer using internal features available to the system before receiving the correct decision from the outside. Therefore, the recording order is part of the model and not a technical detail.

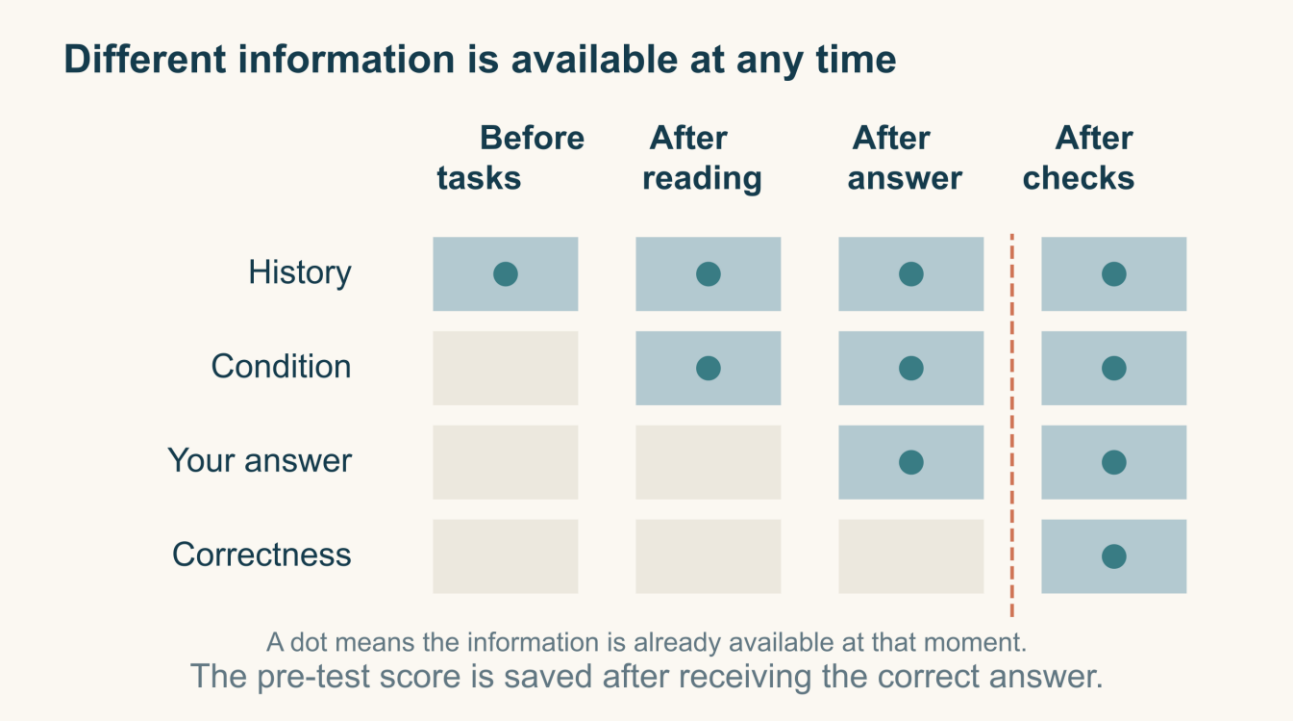

Figure 26. Information before and after feedback. The history of past results, the condition of the current task, your own answer and its correctness become available sequentially. Moment separation preserves the distinction between a prediction before feedback and an estimate with a known solution.

Fleming and Lau emphasize that different measures of metacognition answer different questions about the relationship between confidence and success. [6] Here we will start with a narrow task: we will learn to estimate the future reliability of answers based on past results. This is not yet a complete model of experienced confidence, nor is it a specific assessment of metacognitive effectiveness.

Worked example of a learning assessment

Suppose the system solves a series of tasks and after each one gets correct: 1 means success, 0 means error. Let's introduce two counters - successful and erroneous responses. Initially, we will put one in each. This is conditional smoothing: before real data, the system does not start with complete confidence in success or failure.

Before each task, the prediction equals the success count divided by the sum of both counters. Initially, this is one half. After the first success, the counters are two

and one, so the next prediction is two thirds. After the second success, it becomes three quarters. If an error then occurs, the success count remains three while the total becomes five: the new prediction is three fifths.

Here are the six learning outcomes created for the chapter: success, success, failure, success, failure, failure. Below, the program first calculates and saves the forecast, then gets the next outcome and only after that updates the counters. The loss variable accumulates the forecast error: the actual outcome is subtracted from the probability, and the difference is squared.

```
outcomes = [1, 1, 0, 1, 0, 0]
successes = 1.0
failures = 1.0
loss = 0.0
for outcome in outcomes:
    probability = successes / (successes + failures)
    loss += (probability - outcome) ** 2
    print(round(probability, 3), outcome)
    successes += outcome
    failures += 1 - outcome
print("Average error", round(loss / len(outcomes), 3))
```

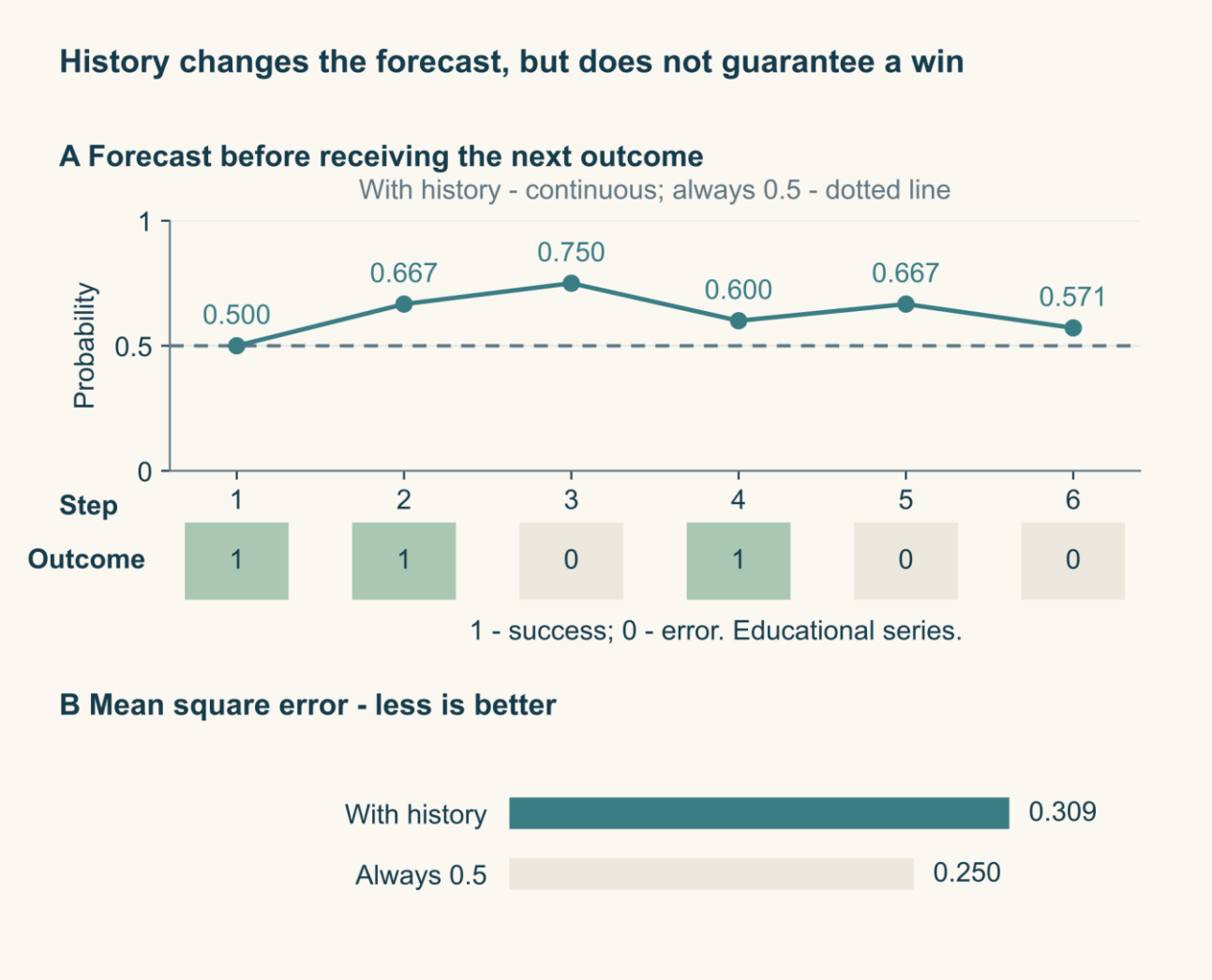

Figure 27. Predictions and error of two models. Before the six training outcomes 1, 1, 0, 1, 0, 0, the model with smoothed counters predicts 0.500; 0.667; 0.750; 0.600; 0.667; 0.571. The dotted line shows the constant forecast 0.5. In this series, the mean square error of the model with a history of about 0.309, a constant - 0.250; this is a series-specific comparison, not a universal superiority.

Pre-assignment grades will be approximately 0.500; 0.667; 0.750; 0.600; 0.667; 0.571. After the entire series, the counters will be equal again, and the forecast for the next task will become 0.5. The change is clear from the mechanism: successes increase one part of the attitude, mistakes - another.

Why comparison with a simple model is necessary

Let's take an opponent who always predicts the probability 0.5. For any binary outcome, its squared error is 0.25. The learning model has an average error in our short run of approximately 0.309. This means that on these training data the constant estimate turned out to be better according to the chosen rule. There is no need to hide the result: it shows that updating based on history is not automatically useful.

The first successes raised the expectation of the model, after which errors appeared. Confidently incorrect forecasts received a large penalty. The six outcomes do not establish the overall superiority of the constant model. To compare rules, we need new series with predetermined conditions: constant success, varying difficulty, or differences between task types.

The squared error evaluates the quality of a probabilistic forecast. It does not measure the depth of experience and does not by itself indicate how well confidence distinguishes between specific right and wrong decisions. The name of the indicator must correspond to the calculation performed. This rule helps to preserve the meaning even with further complication of the model.

Overall reliability and difficulty of this case

Our counters treat tasks as instances of the same type. However, the actual series may include easy and difficult questions. The same general forecast for each subsequent task then loses useful information. If the type is known in advance, separate counters can be maintained by type. If difficulty is assessed by the internal ambiguity of a solution, the model must indicate which feature is used for this.

For example, a classifier may obtain similar scores between two options. A slight difference between them is a possible sign of uncertainty, but its significance must be checked against the correct answers. High internal numbers do not become good confidence just because of the name. The model should show the relationship between such attributes and actual success on new cases.

This is how we distinguish between the general prediction “I usually solve problems like this” and the assessment “this particular answer is reliable.” The first uses the history of a class of cases, the second can take into account information from the current decision. A good research project compares them and finds out what each source of information adds.

When the past is no longer equally useful

Let's say the tasks gradually become more complex. A long history of easy tasks may support an overly high prediction. One way to adapt is to reduce the weight of old outcomes before adding new ones. Another is to maintain separate regimes and try to determine which one is in effect now. Both options require assumptions about how the environment is changing.

Rapid forgetting helps respond to change, but makes assessment sensitive to a random series of errors. A slow update is more resistant to short-term noise, but may lag behind the actual change. This is where a testable trade-off comes into

play. You cannot choose a universally best tempo just because the system wants to “be flexible”; you need a set of conditions that represents the task of use.

Automatic update of counters is a mechanism for regulating assessment. Explicit reflection involves an additional step: considering the reasons for the rule itself. For example: “My confidence is based on overall success, although the tasks have become of a different type.” You can then propose a new assessment organization and compare it with the old one. When designing a system, these functions should be described separately.

Why is a probabilistic forecast penalized?

Before the third outcome of the training series, after two successes, the model predicts 0.75. The third outcome is zero, so the square of the difference is 0.5625. The constant forecast 0.5 receives in the same sample 0.25. This is not a learning penalty per se: the rule penalizes confident errors more strongly. If you first include the third outcome in the counters, and then call the new number a prediction of the third outcome, the estimate will use the answer. This order checks another property and creates a data leak.

What has now been established

We have determined when a forecast is made and against what it is evaluated. History can help or hinder. Its benefit is the result of comparison, and the name of the mechanism does not guarantee a win.

Practice 12 Check to continue

Why do it: Compare constant and learning forecasts on continuation, without selecting a rule based on the outcomes being tested.

What you will need: Success rates and distinguishing between currently and future information. The squared error below is used as a pre-selected forecast evaluation rule.

Time reference: 25–35 minutes. Can be done intermittently; This is not a speed standard.

1. Start with success and error counters of one. For outcomes 1, 1, 0, 1, 0, 0, write down the probability before each update.
2. For each prediction, calculate the square of the difference with the current outcome. Compare the average with the constant forecast 0.5.
3. Keeping the final counters, add a continuation 1, 0, 1, 1, 0, 1. Compare the second part separately.

4. Before a new change, select a forgetting rule if you want to explore it. Check the parameter selected as a continuation on another series.

What to keep: Temporary table of forecasts, two separate comparisons, and justification for the order of calculations.

Hint: After the first series, the counters are four and four. The forecast for the next outcome is again 0.5.

Research continuation: Divide tasks into known types. Compare the general counters with the individual ones, and then formulate what data will allow you to notice a change in regime.

Analysis and answers - Appendix F, chapter 12, p. 285

Self-test questions

1. Why can't a prediction after receiving the correct answer be assessed as initial confidence?
2. Why is the cumulative percentage of successes not sufficient to evaluate each individual task?
3. How does automatic updating of counters differ from explicit reflection on the evaluation method?

We now have a language of state, feedback, source, and evaluation. It will allow theories of consciousness to be read as proposals about structure and predictions, rather than as a list of names.

Part 4

Theories and cutting edge research

Chapter 13

Workspace and recurrent processing

In previous chapters, we separately examined state, feedback, content use, and response evaluation. These concepts will now allow us to compare two theoretical ideas: refinement within sensory processing and dissemination of the result to different consumers. Let us first trace their functions in a fully specified device. Then we return to the stronger claims of the theories of conscious access and experience. The coincidence of functions does not make two theories identical.

From name to device

Let's imagine that you are reading a message and a cup falls nearby. The sound interrupts the reading; you turn, assess the danger and change your action. In explaining such an episode, it is easy to combine several different issues. How was the sound recognized? Why did it get the advantage? How did information become available to movement, memory, and reasoning? Why was the event experienced? The theories discussed in this chapter offer different ways to connect these issues. Comparing them requires first understanding the proposed mechanisms and then separately assessing the transition from mechanism to consciousness.

It is convenient to think of a workspace as an organization of exchange between specialized processes. Recurrent processing means that the result of processing returns and changes the further operation of elements that have already participated. These are different characteristics of the device: the breadth of availability and the presence of reverse interaction. They can be combined. Therefore, the debate between approaches cannot be reduced to a choice between a “connected” brain and an “unconnected” brain.

The Central Claim of Global Workspace Theory

In the global neural workspace model of Stanislas Dehaene, Michel Kerszberg and Jean Pierre Changeux, specialized processes interact with a distributed system capable of maintaining and disseminating selected information. The computer implementation of 1998 linked its participation with mastering the task, requiring

effortful execution and error handling. This is a specific proposal about the organization of cognitive work, not the discovery of a separate brain "screen". [16]

Let's look at our own teaching example. The device has four modules: object recognition, risk assessment, movement selection and event memory. As long as the recognizer sees the cup, only it stores the corresponding record. After selection, the “cup falls” message receives a common event identifier and is transmitted to other modules. The risk assessment associates it with a possible burn, the movement chooses to retreat, the memory retains the incident. The architecture explains the coordinated use of the same content by multiple processes.

Now let's clarify the statement being verified. If a common channel is needed to transfer new information between modules, its temporary shutdown should disrupt the transfer. Recognition of a cup may persist, but the use of the recognized one in a new task may disappear. If the shutdown simultaneously impairs sensory input and movement, the results are mixed. We will not know which of the damaged processes was responsible for the change in behavior.

The strong version of the conscious access explanation attributes a special role to the transition to sustained widespread availability of content. It is the responsibility of the researcher to indicate what is considered sufficiently widespread, what process supports the content, and what data would distinguish such a transition from normal signal amplification. The expression “information has become global” without these clarifications remains the name of the result.

The Central Claim of Recurrent Processing Theory

In Victor Lamme's approach, the initial forward propagation of the visual signal is separated from subsequent feedback interactions. Local recurrent processing is considered a candidate mechanism for visual experience, while broader interactions are associated with access and report. A strong thesis is the possibility of an experience that has not yet received all the possibilities of subsequent use. This is a theoretical interpretation that requires independent verification. [17]

Let's build another learning mechanism. The first module highlights short lines, the second assumes the outline of an object. The assumption goes back to the first and changes the weight of the ambiguous lines. After several exchanges, the system selects the agreed form. Here the feedback makes clear the dependence of recognition of the part on the whole. However, the presence of a cycle does not yet explain why processing should be accompanied by the visibility of the form. A fan with a temperature controller also forms a control loop; its structure differs significantly from the described visual interaction.

Therefore, the recurrent hypothesis must be written with restrictions. What elements are exchanged between? What is sent back? How long is the interaction? How to distinguish meaningful feedback from general modulation of excitability? What operations should be preserved if it is violated? The more detailed the answer, the less opportunity there is to declare any cycle found as a confirmation.

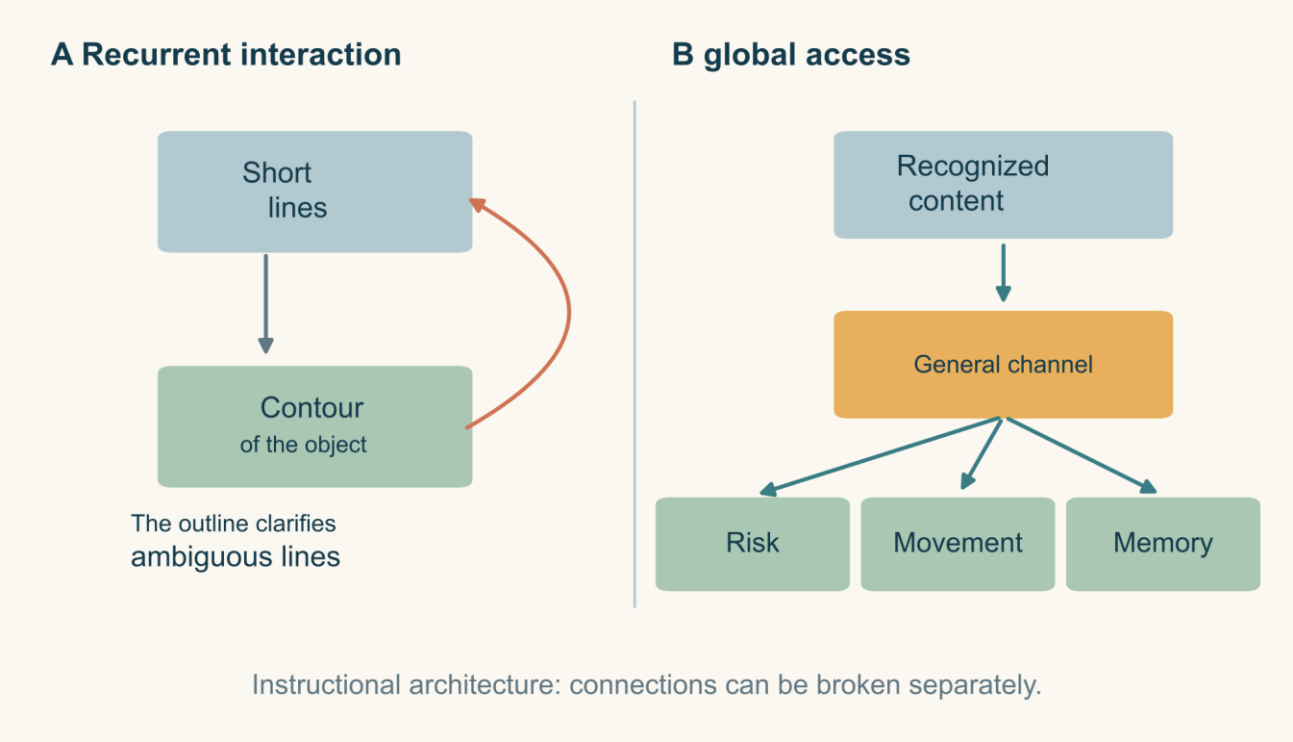

Figure 28. Recurrence and shared access. Two aspects of one educational device. Returning from the contour to the features helps to coordinate recognition; a shared channel allows multiple processes to use selected content. These connections can be changed separately. The diagram is not a test for the presence of consciousness.

In the study by Fahrenfort, Scholte, and Lamme, visual masking was combined with preservation of early processing and disruption of later signals interpreted as recurrent. The work supported the distinction between processing stages. It does not establish that any late response is sufficient for the experience, nor does it exclude all additional conditions of conscious perception. [18]

From a Diagram to a Theory's Commitments

It is useful to go through one case twice. First explain the processing: local features provide several possible objects, the context limits the choice, the result becomes available for further action. Then ask about the conscious status of each stage. The first pass sets the mechanism; the second shows what additional point of the theory connects the stage with the experience.

For the workspace approach, what is essential is not the arbitrary copying of a number, but a certain selection, maintenance and widespread use of content. You

need to list the consumers, the new uses, and the consequences of selectively limiting access. Therefore, simply passing to one preconfigured output is a preliminary example, not the whole theory.

For the recurrent processing approach, it is important to clarify which feedback interactions are assumed to be sufficient for visual experience and how they differ from the subsequent report. The final sequence of repeated calculations can be expanded into a direct circuit. This forces a distinction between computational function and physical organization if the theory gives importance to the latter.

A good comparison selects one divergent forecast. Let's say that local content is preserved, but has become inaccessible to an arbitrary response. The dispute then concerns what basis allows us to consider a persisting process to be an experience and what independent indicator confirms this. The dispute cannot be resolved by simply renaming the process unconscious or conscious. The version being reviewed, the measurement, and the revision condition must be agreed upon.

The training system before practice is intentionally simpler. Its task is to provide experience in disabling functions separately. Once completed, we can return to the scientific debate with a precise understanding of which parts of the argument have been verified by such a calculation and which require additional data.

Where is the real disagreement?

Let us imagine content that is stably represented in the visual system, but is not used for voluntary response. One researcher will say: the experience is already there, access is limited. Another will answer: processing that precedes conscious access has been found. The disagreement concerns which device is considered sufficient and how to establish the presence of the sought-after phenomenon, regardless of the accepted theory.

This difficulty does not make the study pointless. Narrower predictions can be compared. For example, one model requires a long exchange between two sensory areas; the other requires connecting an additional system during a certain period. An experiment that selectively changes one interaction while maintaining the other is then useful. If an intervention is indiscriminate, it does not discriminate patterns well even when the behavioral effect is strong.

In our training device, you can independently turn off the return from the loop to the lines and the general channel of the event. The first time you disable it, form coordination will suffer; the second time, its use by different tasks will suffer. These differing consequences indicate that the two mechanisms serve different functions. To move to a claim about consciousness requires additional

argumentation. The device is currently giving us an experiment with information processing.

You should not expect in advance that one process will replace all others. A mechanism may be necessary in one environment and make do with another mechanism after training. The detection threshold depends on the strength of the signal, and the willingness to respond depends on the instructions and the cost of the error. Therefore, the change in outcome must be decomposed into changes in sensitivity, strategy, response time, and subjective assessment.

Let's add a second source to the training system - a message about a moving person. Now the general channel must choose between two contents. If the cup always gains an advantage, the device will cope with one example, but will not show flexible selection. Let's change the task: sometimes you need to remember a person, sometimes you need to prevent a collision with a cup. Before checking, we will specify how the instruction affects the priority. Then we'll see if the recognition of the losing content persists and can later be accessed by other modules.

This complication no longer checks the presence of a communication line, but the rules of its use. It helps separate a rigid command sequence from an architecture that can convey different content depending on the new purpose. Both devices can perform the initial test successfully, but their differences will become apparent when transferred to an unfamiliar combination of requirements.

One sample with clarification and transfer

In the new training system, the local attribute is equal to 0, 1 or a question mark. The context contains 0 or 1. If the feature is defined, the recognizer saves it. If it is undefined, initially writes 0; enabled context return replaces this entry with the context value. The general channel copies the final content into memory. Without transfer, the memory remains 0. In the sample (?, 1), the full system first receives the local 0, then refines it to 1 and transfers 1 to memory. Disabling the return leaves 0; disabling transmission retains recognition of 1, but the memory remains 0.

The correct label for a specific input is equal to that input for an undefined context. The appraiser needs it. Here we examine the operation of predefined rules. A competitor with direct access to all inputs can solve the same problem, so the success of the demonstration does not establish the uniqueness of the workspace or the adequacy of the return for the experience.

What has now been established

We obtained two independently variable relationships: clarification of content and transfer of it to memory. Now the four modes can be compared in terms of two results, without calling the loss of one of them the disappearance of consciousness.

Separate functions of two links

Mode	Recognition	Result memory
Full scheme	10 out of 10	10 out of 10
Without feedback	7 out of 10	7 out of 10
Without common channel	10 out of 10	4 out of 10
Without both connections	7 out of 10	4 out of 10
Direct copy	10 out of 10	10 out of 10

10 fixed inputs from the analyzed example

Equal output of the two circuits requires the following discriminative experiment.

Figure 29. Independent shutdown of two mechanisms. The result of ten fully specified training trials; reproduced teaching_lab.py --chapter 13. The return clarifies the ambiguous sign, the common channel transmits the memory response. Direct copying is a strong functional control. The table does not measure consciousness and does not represent a comparison of complete scientific theories.

Practice 13 Translation of theory into prediction

Why do it: Test separately the contribution of refinement and transfer on identical ten samples.

What you will need: Input, status and output; difference between local representation and its use. Before practice, all the rules of the training system and ten inputs are given.

Time reference: 35–45 minutes. Can be done intermittently; This is not a speed standard.

1. Take pairs of local attribute and context: (0,0), (1,0), (0,1), (1,1), (1,0), (?,1), (?,0), (?,1), (?,0), (?,1). For each, calculate the correct label using the rule above.
2. For a complete system, fill in the columns: initial recognition, final recognition, memory content. After each trial, the memory begins with 0.
3. Repeat the same ten entries without return, without transmission and without both mechanisms. Calculate the accuracy of recognition and memory separately.
4. Check the table with the command python teaching_lab.py --chapter 13. Then add a direct competitor that sees both inputs and refine the limits of inference.

What to Keep: Four modes, two results for each, and an explanation for one specific discrepancy.

Hint: Disabling transmission does not change the recognition rule. If both indicators change, check to see if you have turned off multiple functions at the same time.

Exploratory continuation: Change the task: after a delay, a new instruction arrives to one of several consumers. Determine the memory, budget and inputs of a strong competitor in advance. Print the result where the mechanisms actually diverge.

Analysis and answers - Appendix F, chapter 13, p. 285

Self-test questions

1. Why aren't global propagation and recurrence strictly opposed?
2. Does the disappearance of the report after the intervention prove the disappearance of the experience?
3. Why compare mechanisms with the same computing budget?

The difference in functions became obvious. The next chapter will add two other topics of theory: internal causal organization and the system's representation of its own state.

Chapter 14

Integrated information and higher order theories

Workspace and recurrence can be compared as ways of organizing processing. However, theories of consciousness also ask what arrangement has meaning independent of the habitual external response and how the state is represented to the system itself. Consider causal structure and higher order representations. Our challenge is to accurately understand the underlying commitments of each theory; the computation of the full IIT formalism will not be silently replaced by the drawing of a few connected nodes.

Two directions of explanation

Imagine two systems that answer the question about the red circle in the same way. In the first, many parts mutually change each other’s state. The second has a table of answers to all resolved questions prepared in advance. Externally, they are difficult to distinguish if the questions are limited. Should what we attribute internally to these systems coincide? The answer depends on which properties the theory takes to be essential: causal organization, representation of one's own state, agency, or some combination of requirements.

In this chapter we will look at two areas. Integrated information theory begins with the assumed general properties of experience and looks for the corresponding physical organization. Higher-order theories relate the consciousness of a state to the way it is represented to the system itself. Both programs offer more than convenient behavior descriptions. That is why their underlying assumptions should be considered as carefully as mathematical models and experiments.

What does integrated information theory say?

In version 4.0, the theory of integrated information links experience to a certain indecomposable structure of cause and effect within a physical substrate. The authors proceed from the existence of experience and highlight its internal belonging, certainty, unity, concreteness and structure. Then they propose physical postulates and formalism. The identity of experience and the unfolded causal structure is the main explanatory claim of the theory. It is not transformed into an established empirical fact by a single mathematical notation. [19]

Let's look at the difference between registering a connection and checking the cause using our own example. Two indicators light up simultaneously every

second. The observer notices a perfect match. However, the cause may be a shared timer; one indicator does not affect the other. If intervention in the first changes the future of the second, and intervention in the second changes the first, the device has a different causal organization. Identical observed sequences do not yet determine internal dependencies.

This example does not calculate integrated information. It teaches us to ask the question prior to computation: what states and interventions do we consider physically possible? Let's say our model has two elements with two states. It is necessary to set a rule for changing each element for all combinations of inputs. If some combinations are prohibited, you must explain why. If two elements actually indicate a common source, treating them separately may distort the analysis.

The next question concerns the system boundary. We can study one module, several interacting modules, or a device together with an external controller. The results of causal analysis may vary. Therefore, the choice of boundary cannot be justified by the desired answer “there must be consciousness here.” First, a selection rule is specified, then it is applied sequentially to the systems of interest and control.

The strength of this direction is the requirement to explain the unity and differences of possible states through the device. Its difficulty for the researcher is the transition from causal organization to the presence of experience. Even an error-free calculated characteristic confirms, first of all, the properties of the adopted model. The connection of these properties with experience must withstand separate comparison with data and alternatives.

What does higher order mean?

Higher-order theories propose to distinguish a state representing the world from a state representing the presence or character of the former state. In Haquan Lau's Bayesian model, conscious evaluation depends on decisions about one's own internal cues and their reliability. This is one specific implementation of a family of approaches, rather than a definition of all its variants. It can explore the discrepancy between task performance and subjective perceptual assessment. [20]

Let's build a training example with a sound recognizer. The first process chooses between two signals. The second receives characteristics of the first's performance: the magnitude of the difference between the alternatives, the consistency of several measurements, and information about recent errors. It evaluates how reliable the current choice is. The third process decides whether to respond now or request another observation. Such a device explains the confidence check if the second process actually uses information about the first.

Let's add an important difference. The variable “confidence equal to eighty percent” may be a simple recalculation of one sensory number. Then it does not indicate a rich model of one’s own capabilities. To show more meaningful control, it is useful to change the sensor robustness by maintaining separate input values. If the system takes into account error history and changes strategy, its score reflects more than the current signal strength.

But functional control of belief does not by itself resolve the issue of phenomenality. A strong theoretical claim must specify which representation of one's own state is sufficient for consciousness and why. It is not reasonable to automatically consider every numerical description of a process as awareness. In addition, conscious perception and detailed reflection are different: seeing color does not mean simultaneously reasoning about the reliability of one’s own color vision.

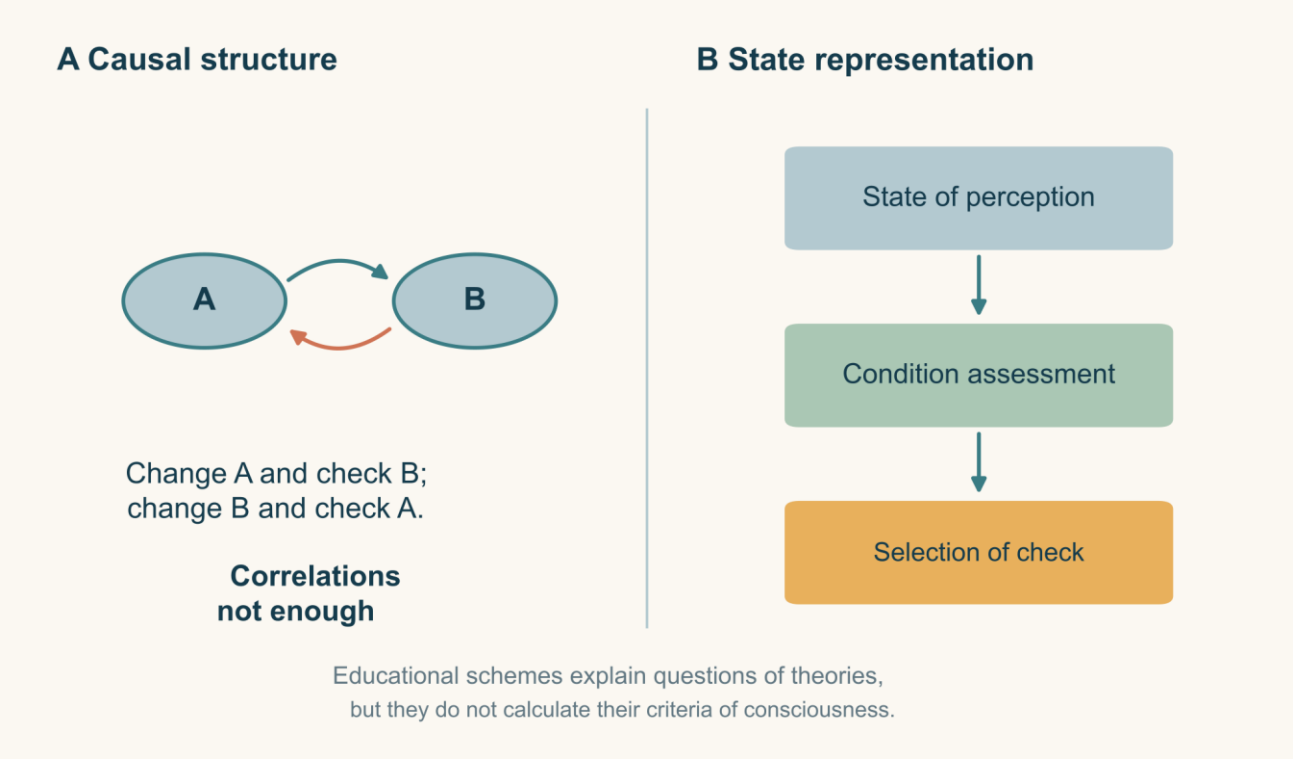

Figure 30. Two different objects of explanation. The left diagram directs attention to the mutual causal influences of the parts; the right one - on the system’s representation of its own state and its use. The diagrams explain various issues of IIT and higher order approaches, but do not replace the formalisms of these theories or assign a winner.

What you need to know about implementation

Two devices may have the same pattern of normal responses and different responses to internal interference. In one, the output depends on the interaction of the states now; another has a pre-recorded sequence. With a limited set of external

issues, they may coincide. Therefore, the analysis of the causal device adds information that is not found in ordinary dialogue.

But any backward arrows are not yet a calculation of integrated information. For a specific version of the formalism, certain elements, states, transitions and a rule for choosing the system under consideration will be needed. The statement of correspondence to experience belongs to the theory itself and is evaluated together with its arguments. The teaching drawing explains the subject matter while maintaining this distinction.

A higher order representation requires a different specification. What is represented as the primary state? How does the second process get information about it? What information is relevant to the current decision itself, and what information is the general expectation of success? Who uses the assessment? These questions allow you to distinguish real monitoring from an external log or a pre-prepared phrase.

Finally, internal access does not have to give complete transparency to the system itself. A process can have privileged access to one signal and be mistaken about another. Therefore, we will check access to specific content, the moment of receipt and causal use, avoiding the general conclusion of perfect self-knowledge.

How to Compare Assumptions

Consider a fictional device with rich internal causal coherence that does not form a separate assessment of its own perceptions. Then imagine a device with detailed confidence control implemented by a different causal organization. Such cases force different approaches to clarify their boundaries. One explanation will ask primarily about physical structure, the other about the relationship between representations.

You cannot honestly resolve a dispute by simply choosing a more attractive example. It is required to indicate how we independently know about the consciousness of the compared systems. If there is no such knowledge, the thought experiment reveals the difference between theories, but does not determine the winner. Its usefulness is in developing testable consequences and revealing assumptions that were previously hidden.

It is useful to distinguish three types of statements. The definition states how the term is used. The formal theorem shows what follows from the rules of the model. An empirical statement tells how the world under study works. From a strict theorem it is impossible to obtain the truth of the initial premises. At the same time, a dispute about premises does not negate the value of formal analysis: it shows the exact cost of the accepted explanation.

For your own research, choose the narrowest conclusion. If the model correctly predicts the request for additional information, this is a result about behavioral regulation. If a causal characteristic distinguishes two networks, this is a result about their organization. The next step towards consciousness must have its own justification; it cannot be discreetly included in the name of the indicator.

It is also useful to check the dependence of the output on the observer. Let's imagine that the developer added a detailed log of internal states that is not readable by anyone inside the device. The researcher now knows more about the system, but its capabilities have not changed. Therefore, the information available to us about a process is different from the information that the process itself uses to control it. If the log is turned into a control module input, the organization actually changes and the consequences can be verified.

This distinction is especially important for self-description. Information recorded by an external experimenter in the first person does not become an internal model of the device. It is necessary to show who uses the recording, what actions depend on it, and what will happen if it is selectively distorted. Then the dispute about the representation of one's own state receives a concrete object.

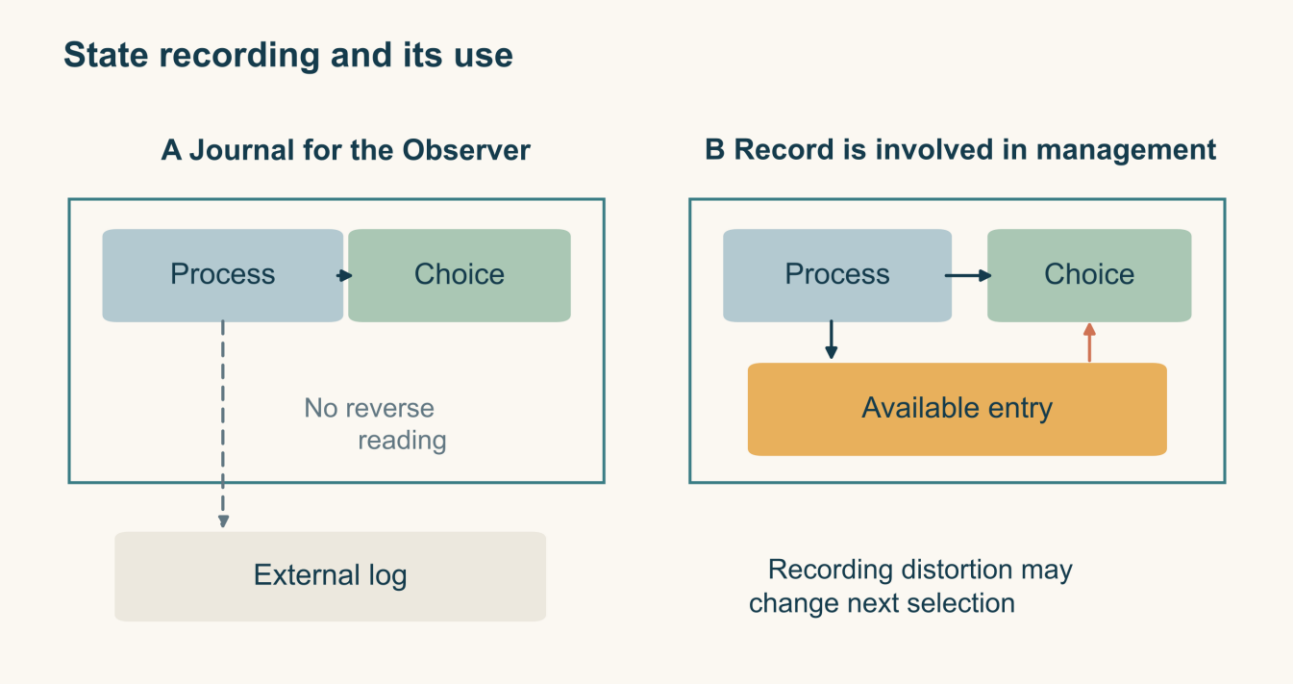

Figure 31. External log and internal use. In the left design, the researcher receives information about the process, but the device does not read the log. On the right, the record becomes a control input: its selective distortion makes it possible to test the causal role. Such testing concerns the use of information and does not yet establish a higher-order representation sufficient for consciousness.

How not to mix definition and strong assumption

The phrase “let’s call an internal assessment a forecast of one’s own error” sets a working definition. The statement “if a report reads this score, changing it can change the report” requires specific model connections. The phrase “appropriate internal evaluation makes a state conscious” already adds theoretical explanation. Finally, the observation of a relationship in a particular sample is related to the data. These sentences can participate in the same argument, but do not receive equal force due to proximity. It is especially important to maintain this analysis when comparing IIT and higher-order theories: they differ not only in metrics, but also in the initial objects of explanation.

What has now been established

General connectivity and the confidence number remain functional properties until a more precise theoretical requirement is formulated. Now let’s try to explicitly write down where the description of the device ends and the assumption about the experience begins.

Practice 14 Assumptions specification

Why do it: Create a specification of two explanations with distinguishable statuses of statements.

What you will need: Distinguish between a mechanism, its measurement, and a statement about its relationship to experience. Necessity and sufficiency are introduced in chapter 5.

Time reference: 30–40 minutes. Can be done intermittently; This is not a speed standard.

1. For a causal system, identify the two states that change and the direction of each influence. For a rating system, indicate the primary decision, the internal reliability indicator, and the recipient of the rating.
2. Write down four points for each approach: the phenomenon being explained, the proposed device, the basis for its significance, and the condition for revision.
3. Label each statement as a definition, assumption, consequence of a given model, or empirical statement. Separate mixed positions.
4. Select one intervention and a simpler mechanism that can produce the same response. If there is no discriminating result yet, save this as a summary.

What to keep: Two passports and one causal test with a clear boundary for phenomenal inference.

Hint: The mutual influence diagram is not an IIT calculation. An error prediction does not become a higher-order representation just by virtue of its name: its content, origin, and use are needed.

Exploratory follow-up: Take a specific version of one theory from the chapter's sources. Specify what its sufficient condition requires beyond the established educational system, and suggest a case where the conclusions of the theories diverge.

Analysis and answers - Appendix F, chapter 14, p. 286

Self-test questions

1. Why is a high correlation between parts not sufficient to infer their mutual causal influence?
2. Is reflection equal to any higher order representation?
3. Can a theory be mathematically precise and empirically controversial?

The next chapter will show a concrete way to calculate hidden causes and one's own reliability through prediction. This will help make some of the relationships discussed feasible.

Chapter 15

Prediction and self-models

The representation of one's own state becomes more meaningful when one knows how it is formed and used. The predictive approach proposes to build versions of hidden causes, compare expected signals with received ones, and update the assessment. Let's look at the world and the consequences of the command separately. Their confusion is especially noticeable after a change in control: the system can clearly see the target and at the same time make mistakes in action.

Why observation requires an assumption

You hear a short sound from behind the door. It could be an object falling, a lock tripping, or a switch being flipped. There is one signal available, but there are several possible causes. To choose an action, you need to take into account not only the signal, but also the context: who is behind the door, what happened before, how well the room can be heard. This example shows the main task of inference: to evaluate hidden causes from incomplete observation.

Prediction in a scientific model does not necessarily mean verbally expecting the future. It could be a calculated guess about what signal should appear given a certain state of the world. Comparing what is expected and what is received helps change the assumption. Rao and Ballard's model used this hierarchical organization to explain some properties of visual neurons: the upper level predicts the activity of the lower one, and the discrepancy is involved in the refinement of the representation. This is a model of processing, not independent evidence of the occurrence of experience. [21]

A simple hidden cause model

Let's look at our own computable example. The target is on the left or right. Before observation, the probability of the right side is estimated at sixty percent. The sensor correctly shows the side in eight cases out of ten. The signal "to the right" appears. How much should the estimate be changed?

Mentally consider one hundred identical situations. At sixty the target is on the right; in about forty-eight of them the sensor will show the right side. At forty the target is on the left; at about eight it will erroneously show the right side. Among the fifty-six right readings, forty-eight correspond to the right target. The updated probability is approximately eighty-six percent. This is a consequence of given numbers and assumptions about sensor reliability, rather than an experimental measurement of human perception.

The example shows why to separate three values: the assumption before the observation, the diagnosticity of the observation and the final assessment. A strong signal may do little to change a belief if the sensor is unreliable. Lack of observation does not equal observation of lack of purpose. If the sensor did not work, the system should maintain uncertainty, taking into account possible changes in the world, and not invent a confirmation of the previous version.

Now let's allow the target to move sometimes. Memory can no longer be understood as the immovable storage of the last answer. The old belief must be changed even before a new observation: the longer the break and the higher the mobility of the goal, the less reason for the previous confidence. This function makes memory part of the current output. In chapter 28, a similar mechanism will be available in a running agent.

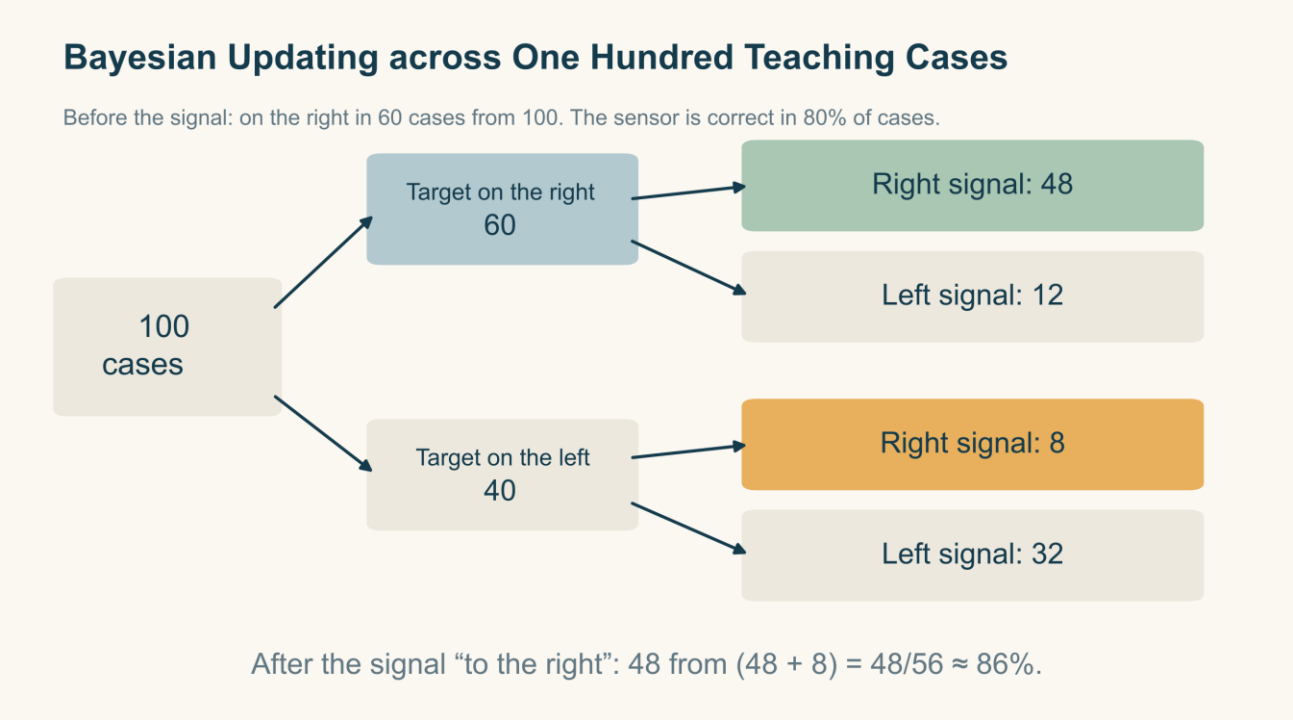

Figure 32. From preliminary estimate to updated probability. The tree uses exactly the numbers of the chapter's case study: the prior probability 0.6 and the symmetric sensor reliability 0.8. With a signal “on the right”, 56 cases remain, of which 48 have a right target. This is a calculation under given premises, and not the observed frequencies of human perception.

From a model of the world to a model of action

Let the device be able to send two commands. As long as the "right" command moves the pointer to the right, knowledge of the target is sufficient for selection. Then the connection changes imperceptibly: the right command gives the left effect. If the device only stores a model of the world, it will continue to know where the target is and at the same time act incorrectly. This means that it is necessary to

separately represent the relationship between one’s own commands and the observed consequences.

In such a device, a functional model of one’s own ability to act appears. Its content is specific: which command is likely to produce which effect. To call this full-fledged human self-awareness would be premature. But it is possible to show exactly what problem a self-referred variable solves: it allows one to distinguish an error in knowledge about the world from an error in controlling an action.

Imagine that the action was unsuccessful. The system must consider at least three reasons: the goal was incorrectly defined, the correspondence between the command and the effect has changed, and the performer accidentally refused. If all errors are automatically attributed to the sensor, learning will spoil good perception. If everyone is assigned to an action, the device will change control endlessly. A good model requires observations that distinguish between causes of error.

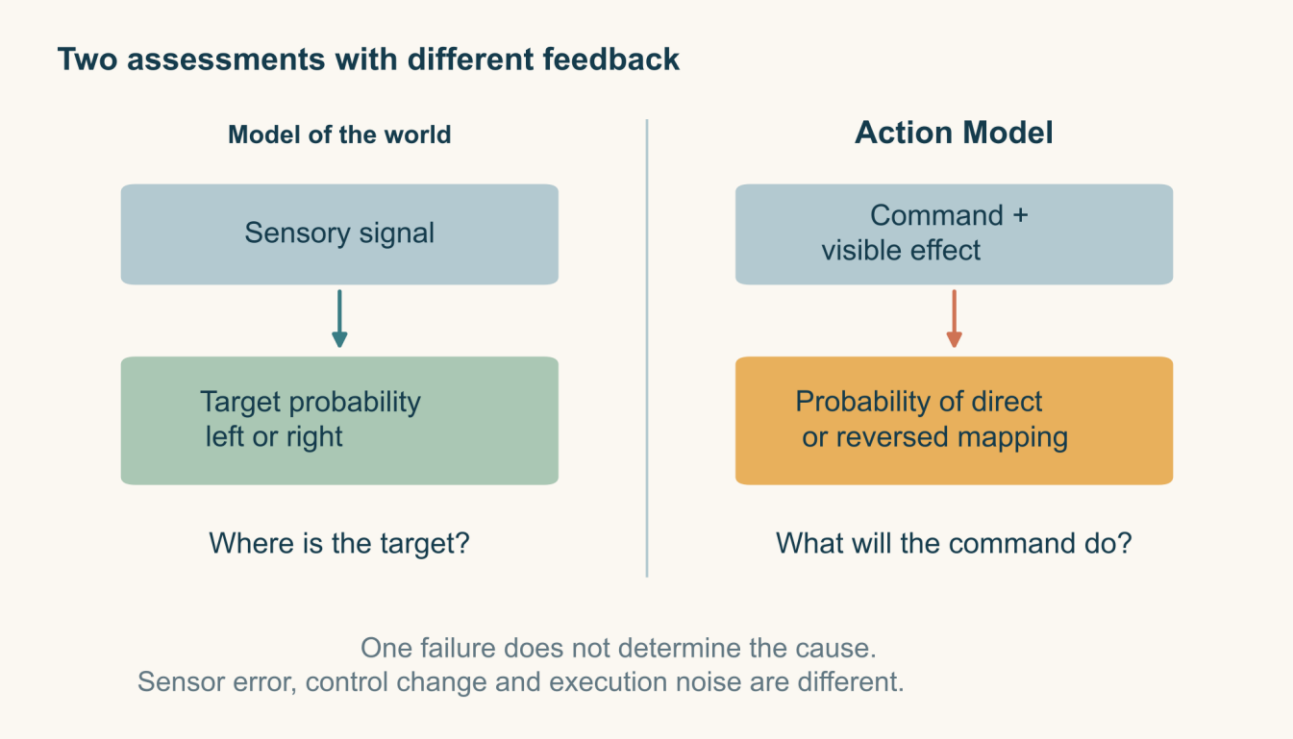

Figure 33. Separate models of the world and actions. The training device stores separate beliefs about the location of the target and the correspondence of the command to the visible effect. Different observations update different estimates. One failure cannot clearly establish a perception error, a change in control, or an accidental failure of execution.

The body as a common cause of signals

In the Bayesian model of Samad, Chung, and Shams, visible hand ownership was associated with common cause inference for visual, tactile, and proprioceptive signals. Consistency in time and space changes the plausibility of their association.

This work suggests a mechanism for certain bodily effects; it does not derive from it all forms of self-awareness and biographical self. [15]

Using your own example, you can understand the meaning of the common cause. You move your finger, see the movement, and receive associated bodily signals. One hypothesis combines them as consequences of one-arm movement. The other involves several independent events. With stable matching, the first model turns out to be more economical in terms of the number of independent matches. However, the conclusion depends on what alternatives are presented at all and what matches the system has learned to expect.

Bodily identity and authorship of an action also do not have to change together. In the experiment of Kalckert and Ehrsson, the conditions of movement of the rubber hand made it possible to distinguish between subjective control of the hand and the feeling of control over its movement. This is an argument against reducing the two assessments to one indicator. It does not mean that in ordinary life they exist in isolation. [4]

Prediction of internal states

Anil Seth proposed to consider emotional states and bodily self-awareness through inference about the causes of signals from within the body. The approach connects prediction with the regulation of bodily state. It is a theoretical program: its usefulness is determined by what specific differences in sensations and behavior it explains and predicts. [22]

To develop your own model, this raises a useful question. The same change in input signal may indicate danger, physical stress, or measurement error. How does the system differentiate between these reasons? In the training device, you can specify the internal resource consumed by the action and the sensor with noise. Then its assessment of its own state will be checked against the forecast of future availability of actions. We will receive a model of resource regulation, without a basis for attributing fatigue to the device as an experience.

The general limit of predictive explanations is the breadth of their application. Many processes can be described with the words "wait" and "error". What creates scientific value is the precise distinction between models: what causes are allowed, how confidence changes, what counts as error, and what action reduces it. If these rules are adjusted anew for each outcome, the explanation becomes too flexible.

Another check concerns the dependence of observations. Let's say two sensors simultaneously show the right side. If they are independently wrong, the coincidence can greatly enhance belief. If the second one simply copies the first one, there is no new information. A model that considers the copy to be

independent confirmation will become overconfident. In a more complex case, both sensors may be affected by a common interference that the device is not yet aware of.

Therefore, confidence control requires a model of the sources of error, not just an update formula. For a practice test, you can compare two identical series of readings: one obtained by independent measurements, the other created by copying. The correct estimate must depend on which procedure generated the series. This example helps to understand why multiple consistent inputs do not guarantee reliable knowledge about the world or one's own state.

One observation and two different uncertainties

Before the signal, the probability of the target on the right is 0.6. The sensor reports the correct side with probability 0.8 and is symmetrically wrong. After the signal on the right, the weight of the right target version is equal to 0.6 multiplied by 0.8, that is, 0.48. The weight of the left version is equal to 0.4 multiplied by 0.2, that is, 0.08. Normalization gives 0.48 out of 0.56, approximately 0.857. This updates the target's position. To select a command, you also need to know whether the right command causes the right effect. If the link is inverted, the more likely right target requires a left command. Refining the sensor does not correct the faulty control model.

What has now been established

World, control, and sensory reliability can be different sources of error. Successful learning requires information that separates these versions. Now let's see what data allows us to update each part.

Practice 15 Distinguishing three errors

Why do it: Separate the assessment of the target's position from the assessment of the command's relationship to the effect.

What you will need: Probabilistic chapter transition 11. Before the action, observation and history are available; the true goal and actual outcome are related to the environment and the evaluator.

Time reference: 30–40 minutes. Can be done intermittently; This is not a speed standard.

1. Repeat the calculation of the right target for the initial probability 0.6 and the sensor accuracy 0.8. Then get a forecast for the signal on the left.
2. For four cases, the target is on the left or right and the connection is direct or reverse, write out a command that will reach the target without motor noise.

3. In each case, indicate what is learned from sensory observation and what is learned from comparing the command to the visible effect.
4. Open the prepared example python teaching_lab.py --chapter 15. He shows a decision table; hide the true target and relationship columns from the model in question.

What to keep: Two probability updates, four solutions, and a map of feedback sources.

Hint: The likelihood of a signal under a hypothesis and the likelihood of a hypothesis after a signal occupy different places in the calculation.

Exploratory continuation: Conduct a paper experiment of twenty trials with a partner or a closed column of the environment. Specify the change of control in advance and do not inform the selector of the commands. Before the result, write down the forecast, after it, update only those states for which information was received.

Analysis and answers - Appendix F, chapter 15, p. 286

Self-test questions

1. Is every prediction a conscious expectation?
2. How does a model of action differ from a model of the world?
3. Why should the absence of a signal be distinguished from a negative signal?

Predictive mechanisms have become more defined. Now let's return to actual research and ask what data actually differentiate between theoretical explanations.

Chapter 16

Where data differentiates between theories

We are able to describe several mechanisms and obtain their expectations. But the same successful result can be compatible with several theories. To move forward, it is important to find the conditions under which their predictions diverge and understand what the experiment measured. Reported and non-reported studies are particularly illustrative of why theory and procedure should be discussed together.

Why confirmation is not enough

Suppose two models expect some signal to be amplified when an image is visible. The experiment reveals an increase. The result may support both models, but does little to help you choose between them. Discrimination requires a situation in which expectations diverge. For example, the first model predicts the persistence of a signal without a requirement to respond, while the second relates it to response preparation. Then changing the instructions becomes a meaningful intervention in the testing conditions.

Theories of consciousness are particularly sensitive to this problem. One experiment includes recognition, retention, confidence, decision, and movement. If everything changes together, the difference found can be explained in several ways. Research work consists of separating a related set of processes as much as the specific task allows. The name of the brain region by itself does not make this distinction.

Agreeing in Advance on the Test

In the COGITATE project, proponents of competing approaches and the research consortium agreed in advance on the design, differing predictions, and interpretations. In the study published in 2025, 256 participants viewed highly visible stimuli of varying durations; functional magnetic resonance imaging, magnetoencephalography and intracranial electroencephalography were used. The results supported some of the expectations and created difficulties for both theories: among them, the lack of the expected sustained posterior synchronization and the overall expected burst upon stimulus disappearance. Consistent predictions were tested, not all possible versions of theories. [23]

For your own project, the principle of prior commitment is important here. Before receiving data, you need to write down what will be considered support, difficulty

and uninformative results. If any observation after analysis is declared expected, the theory ceases to risk error. This also applies to the textbook author’s favorite hypothesis: personal interest does not replace the conditions for its revision.

Nor does prior registration make the interpretation infallible. You can choose a weak indicator or a too rough prediction in advance. Therefore, two tests are needed: whether the experiment follows the written design and whether the design actually tests the central claim. If the latter is in doubt, the wording should be explicitly changed and a new test should be carried out, keeping the previous result in the history of the study.

The report changes the problem under study

A study by Cohen and colleagues in 2024 compared visual masking in conditions with and without a report. The late component P3b and some robust decoding characteristics were associated with report execution. In the no-report condition, the earlier component N2 corresponded to a nonlinear change in visibility. It is the result of a specific task and measurement method; it does not set a universal time for the appearance of any experience. [24]

When a person needs to report what he has seen, he must convert the perception into a response, store the necessary information and carry out the rule. Therefore, the disappearance of the signal after canceling the report may indicate its connection with these operations. But the absence of a requirement to respond does not guarantee the absence of internal naming, attention, or remembering. The participant is able to continue to perform part of the previous task independently.

The opposite difficulty also arises. If you don't ask about a perception, how do you know that it has arisen? Some designs use indirect measures, such as eye movement patterns. Their connection with experience must first be justified, and then transferability to new conditions must be checked. Otherwise, a method designed to exclude a report simply replaces one controversial indicator with another.

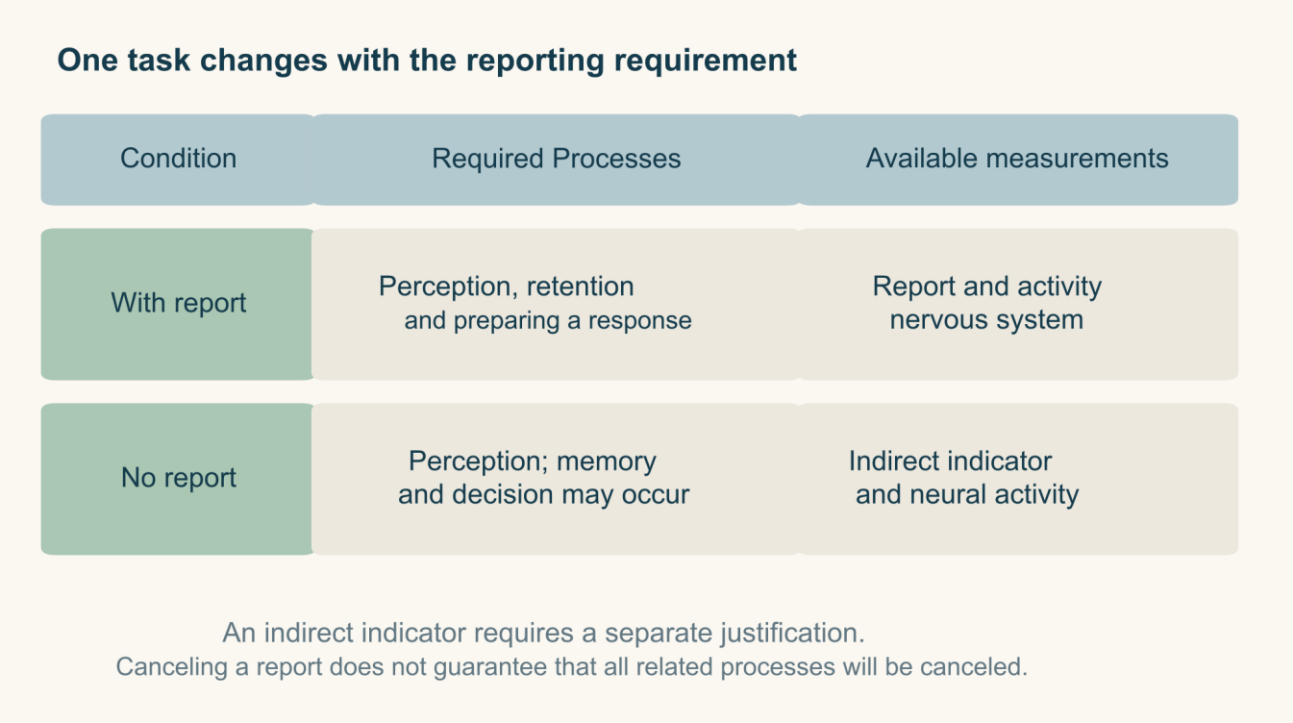

Figure 34. Report and measurement boundaries. Conceptual comparison of conditions with and without explicit reporting. Removing the response requirement changes the task but does not automatically remove attention, memory, or internal decision. Indirect signs of perception also need to be verified. The table does not contain quantitative research results.

Why the Anterior and Posterior Cortex Do Not Form a Simple Choice

He and colleagues, in published work 2026, compared visual perception with and without a report while recording magnetic brain activity. Representations in posterior sensory areas were maintained in both conditions, whereas the detected prefrontal representation of content was dependent on report. This supports the separation of content processing and task demands. Observational recording, however, does not prove the causal sufficiency of a single region. [25]

In another published paper, 2026, Zhang and colleagues found differences between frontal subsystems. Decoding associated with frontal regions of the dorsal attention network carried over between report and no-report conditions; other frontal representations were dependent on report. In the no-report condition, oculomotor measures were used. The result prevents the generalization “all frontal activity relates only to the response,” but does not establish that the selected network itself creates the experience. [26]

It is useful to read these works together, without averaging them into one slogan. They may differ in stimulus, attention, how awareness is measured, spatial precision, and analysis sensitivity. “Could not decode content” means that the selected analysis did not extract it from the data under the specified conditions.

This is different from the claim that content is nowhere represented and plays no causal role.

Successful decoding must be interpreted with equal caution. The classifier may use a perceptual consequence, a common source of variation, or a spurious feature. The mere presence of stimulus information in a signal does not determine the function of this information. The question of necessity requires additional intervention or strong causal argumentation.

How to read contradiction productively

Let's say one article reports a connection between signal and visibility, and another does not find it. Start with the exact subject of comparison. The signal could be measured at different time intervals; participants may have used different strategies; the lack of effect may have been accompanied by too much uncertainty. First, it is determined whether the results are inconsistent under comparable conditions, and only then a revision of the mechanism is discussed.

The following table is useful for educational analysis. It compares the logical strength of results rather than judging publications by prestige.

Observation	Direct conclusion	What has not yet been established
Signal associated with report	Dependence on response conditions detected	That the signal is not involved in other functions
Content is decoded	Data carries discriminating information	What area is necessary or sufficient
Intervention changes outcome	There is a causal effect under these conditions	What exactly has changed is the experience
No effect detected	The analysis did not reveal the expected difference	That the distinction is missing with useful precision

What Changes When a Report Is No Longer Required

Suppose the signal is related to visibility in a task where a response must be made after each presentation. Versions about the perception and preparation of the message are possible. Maintaining a signal under a justified condition without a required response limits the second account, but does not establish the causal sufficiency of the signal for the experience. The disappearance of the signal also requires verification: along with the answer, attention and memory could change. Therefore, one question is formulated narrowly: what characteristic is preserved

when the requirements of the report change and in what independent way the perception itself is confirmed.

What has now been established

The research becomes discriminating against clearly defined competitors. A large sample does not replace the desired comparison, and the absence of a statistical finding does not mean an exact null.

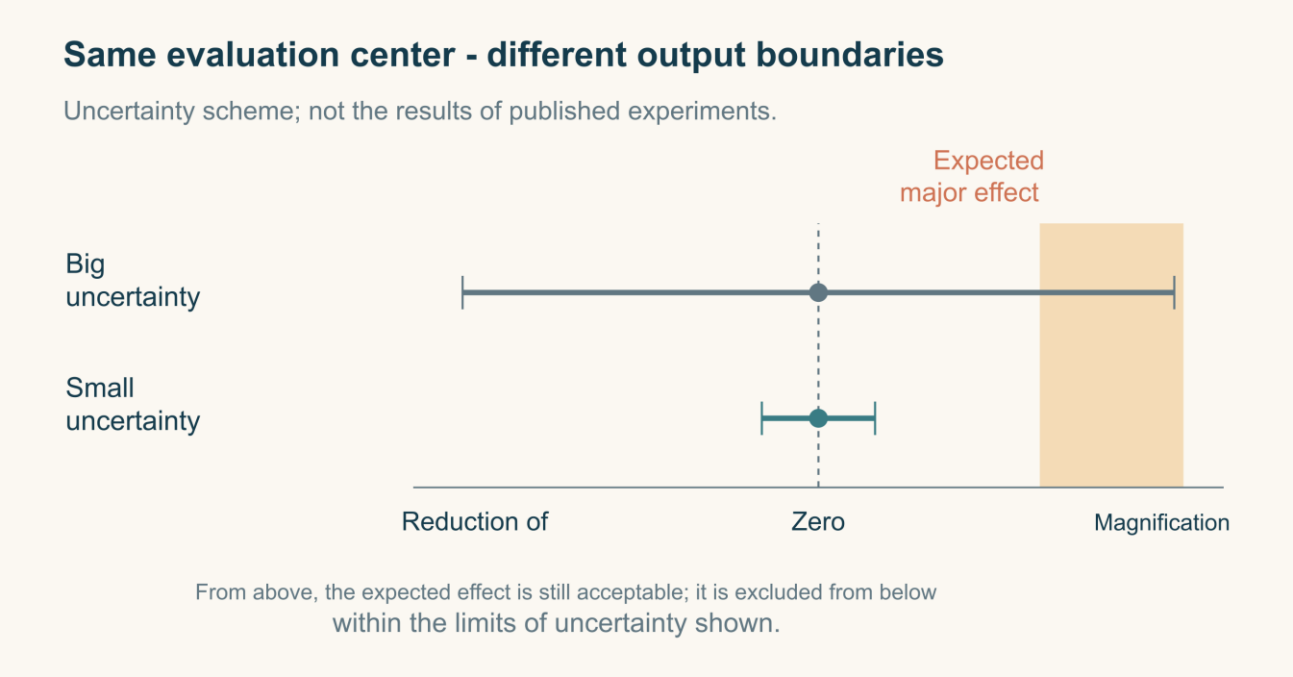

Figure 35. Why lack of significance is not exactly zero. The two schematic intervals have the same center around zero, but allow a different range of effects. Only the narrow interval in the figure excludes the expected large increase. The numerical scale is deliberately not specified: a real test requires an estimate of the effect, a reasonable interval, and a quantitative promise of the theory.

Practice 16 Comparison of two studies

Why do it: Compare two studies on the same question, while maintaining differences in methods and conclusions.

What you will need: Measurement specification from chapter 6, comparison quality from chapter 7, magnitude and uncertainty of the effect from chapter 8. Decoding means extracting discriminative information by a selected analysis.

Time guide: 35–50 minutes for basic analysis. Can be done intermittently; This is not a speed standard.

1. Select two jobs from the chapter with and without report conditions, for example sources 24 and 25. First, use their analysis in the text, without taking credit for reading the full original.

2. For each, fill out five fields: stimulus, task, visibility indicator, measured activity, main comparison. Mark unknown information explicitly.
3. Write down one general question and one supporting and limiting observation for each activity.
4. Make the following prediction when changing the response method and ask whether the competitor expects the same result.

What to keep: Two comparable passports and one prediction indicating whether it distinguishes between theories.

Hint: Do not transfer the result between EEG, MEG and tomography only based on the similarity of words. Separately indicate the time interval and method of analysis.

Research follow-up: Open both primary publications, clarify omissions, and check whether the tasks being compared have changed. Separately estimate the magnitude and uncertainty of the key effect and maintain a distinction between the authors' conclusion and your own.

Analysis and answers - Appendix F, chapter 16, p. 287

Self-test questions

1. Why shouldn't a major study finally settle the theory debate?
2. Is the no-report condition a direct measurement of experience?
3. What changes confidence in a model more—the expected outcome or the discriminative prediction?

The comparison of theories prepared the question of emergence. Let's first narrow it down to one episode: how content emerges and is maintained over time.

Part 5

The emergence of consciousness and self-awareness

Chapter 17

How does a single episode occur?

So far we have compared mechanisms and indicators. Now let us clarify what the appearance of a conscious episode means. This question is different from the development of the child, the evolution of species, and the basis of the very existence of experience. In this chapter, we will analyze only the dynamics of one event: the launch conditions, the maintenance process and the subsequent message. The threshold accumulator from chapter 9 provides a language for such a task, but not a universal time limit for consciousness.

Four Meanings of Emergence

The question “how consciousness arises” combines problems of different scales. The dynamic task concerns a single episode: why a certain content appeared now and what supported it in the next moments. The developmental problem examines the change in capabilities of one organism over the course of life. The evolutionary problem asks about the history of mechanisms in lineages. The ontological problem concerns why a certain physical process is accompanied by experience at all. The answer to one problem can help another, but does not replace it.

This chapter deals with the dynamic issue. For example, let's take the brief appearance of a silhouette in a window. It is necessary to distinguish between the system’s preparedness for perception, the processing of the received signal, the content of the experience and the subsequent response. Even if everything happens quickly, logically these are different links. The latest record of an event need not coincide with the beginning of the event, and the earliest response to light need not be the beginning of conscious vision.

Preconditions Process and Consequences

A premise makes an event possible under given conditions. The process directly implements the change of interest according to the model being tested. A consequence occurs because the event has already occurred. These roles are determined by causal relations, not just by the order of the clock. One process is

capable of participating in several roles: the memory of the previous stimulus influences the next episode, although it itself was formed as a consequence of the previous one.

Let's consider a training setup. A figure appears on the screen; the participant reports its form after a few seconds. The advance instructions tell you what to pay attention to. Sensory processing discriminates shape. Memory holds the result. Selecting an answer converts it into a movement. If memory deteriorates, the later report will become less accurate, even when the initial discrimination has not changed. Therefore, inferring the occurrence of an episode requires monitoring the time chain.

The reverse bias occurs when an early change is considered a sufficient explanation for the entire episode. Let's imagine a sensor that quickly distinguishes between two stimuli. It does not follow from this that the difference is accessible to further behavior or that it is experienced. The name “early marker of consciousness” should be accompanied by an explanation of why the indicator refers specifically to conscious content and not to its input condition.

Instructional dynamics of emergence and maintenance

Let's build our own simple model. It has an activity number from zero to one. At each step, half of the previous activity is retained. The presence of an input signal adds four tenths. If the previous activity has already reached half, an additional internal rule adds three tenths. The value is limited to one. All numbers are chosen for transparency of calculation; they do not assess human brain parameters.

Two consecutive inputs give an activity of four tenths and six tenths. Then the entrance disappears. If there is internal support, activity remains at six-tenths; without it it falls first to three tenths, then to fifteen hundredths. The same entry history gives rise to different durations of the internal state.

Step	Input	Without internal support	With internal support
First	Yes	0.40	0.40
Second	Yes	0.60	0.60
Third	No	0.30	0.60
Fourth	No	0.15	0.60

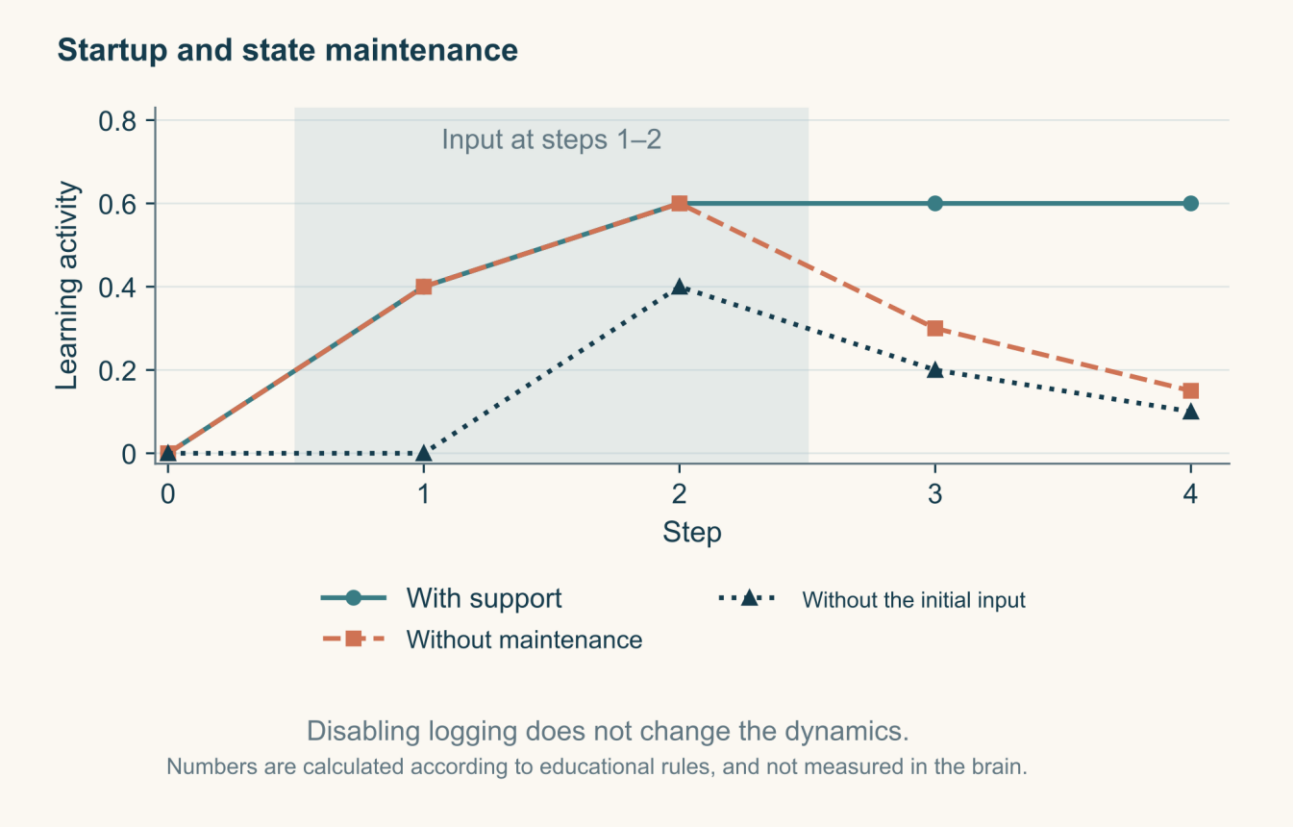

Figure 36. Three trajectories of one learning mechanism. Calculation of the chapter rule 17: at each step, half of the previous activity is retained, the input adds 0.4, and with the previous activity not lower than 0.5, internal maintenance adds 0.3. With two inputs, the supported activity remains at level 0.6. Removing the first entry and deleting the maintenance have different consequences. The observer log entry is not included in the status rule.

In the model, you can examine starting and saving state separately. One input is not sufficient to trigger steady state; two are suitable. After launch, further entry is no longer needed according to the given rules. However, we obtained a stable variable, not evidence of experience. Moreover, the number alone does not explain how the red circle differs from the quiet sound. Meaningful explanation requires representations of the differences between possible episodes.

In a real study, similar logic requires different interventions before, during maintenance, and after the end of presentation. If the intervention only affects triggering, this is one limitation of the model. If only for preservation - something else. The coincidence of effects is also informative: perhaps the proposed division does not correspond to the design of the system or the intervention affects too many processes at once.

What does the change in mechanism show?

Two studies by Noorman et al, using brain electrical activity, examined the effects of memantine on visual figure processing. Changes were found in neural measures interpreted as recurrent integration; Perceptual sensitivity did not improve. The

version of eLife used here was published on June 18 2025. For our question, the gap between a change in the proposed mechanism and the absence of a corresponding change in behavioral sensitivity is significant. [27]

This result teaches us not to consider all indicators interchangeable. Perhaps what is being measured is a process that is not limited to that particular task. It is possible that the behavioral measure is not sensitive enough to the change of interest. It is possible that the initial mechanistic interpretation is incomplete. Choosing between explanations requires new data. It does not happen automatically in favor of your favorite theory.

The rule for your own model follows from this. If turning off a variable does not change the outcome being measured, you cannot immediately conclude that the variable is not needed for anything. It is necessary to check whether the intervention actually worked, whether the task was sensitive to its function, and whether another mechanism compensated for the change. However, the endless addition of unverified compensations also does not save the hypothesis: each must generate a new verifiable consequence.

Content time and message time

When someone says, “I noticed it immediately,” the phrase reports their assessment of the event. Turning it into a temporal measurement requires instructions, a scale, comparison conditions, and an error estimate. The press of a button cannot simply be equated with the onset of an experience: decision-making and movement may intervene. Subtracting the average movement time does not remove every difficulty, because the different stages may interact.

It is useful to seek agreement across multiple metrics with different constraints. For example, changes in visibility at different presentation durations, the nature of errors, subjective assessment, and neural dynamics provide different aspects of the episode. Coincidence does not guarantee a definitive explanation, but it does help rule out some alternatives. The discrepancy should be saved as a result and not resolved by renaming the indicators.

A dynamic explanation must also describe completion and change of content. A model that only knows how to increase activity may get stuck on the first stimulus. It will need a mechanism for termination, competition or renewal. Checking the sequence of different inputs is therefore often more informative than demonstrating one successful recognition.

A separate problem arises when averaging time. Let's say that in each trial the training system switches abruptly, but the moment of switching is slightly different. The average curve across samples will appear smooth. The opposite

situation is also possible: selecting samples based on the threshold can create the impression of a sharper transition than exists in individual trajectories. Therefore, the shape of the average signal does not uniquely determine the dynamics of a single episode.

For verification, save the complete trajectories of individual samples and their sampling rules. Compare two models: a gradual change in each trial and an abrupt change with different start times. If they describe the mean curve in the same way, an additional measure of distribution or transport over time will be required. This is a concrete way of making the question of the occurrence of an episode more precise, without assigning one moment as the universal boundary of consciousness.

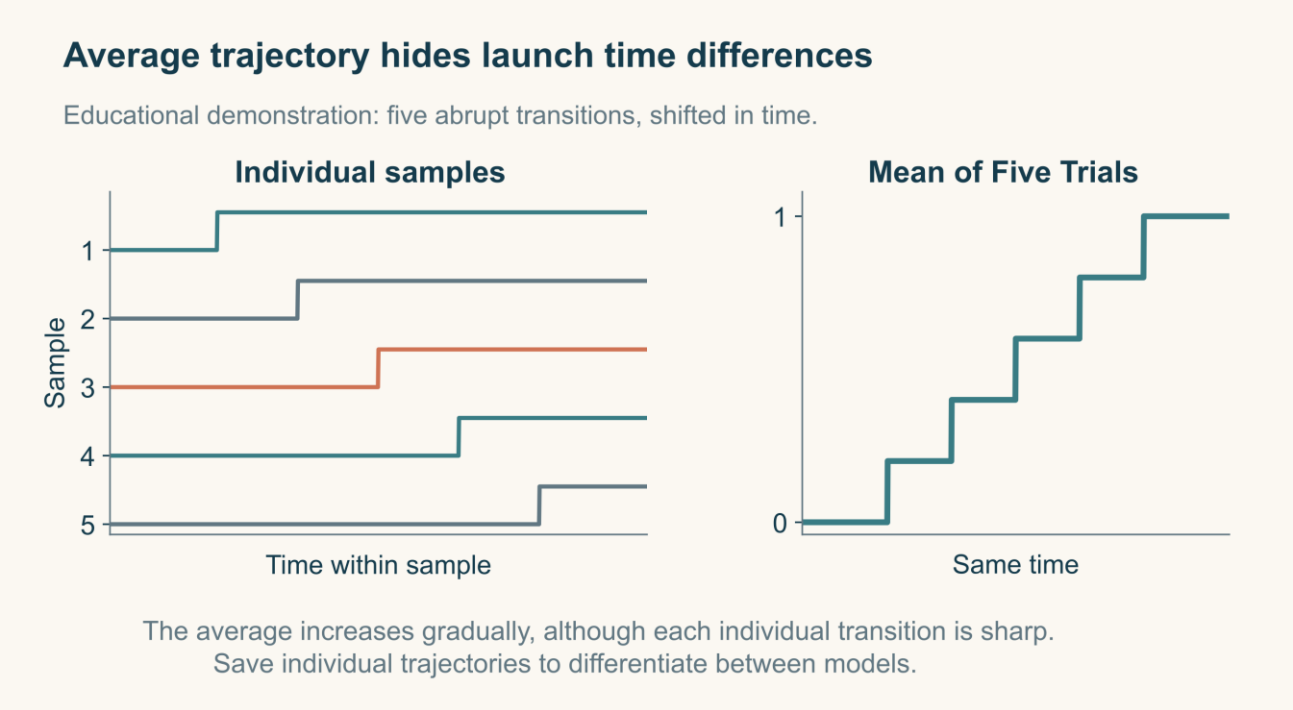

Figure 37. Averaging and dynamics of a single episode. In five artificial samples, the state changes abruptly from zero to one, but the transition moments are different. On the right is their exact average, which increases in steps; with many shifts, the middle curve can appear smooth. This example explains the problem of measuring dynamics and does not indicate the moment of occurrence of the experience.

Why the Mean Curve Does Not Determine the Dynamics of a Single Trial

Five samples may contain the same sharp transition from zero to one, but are triggered at different moments. The time average then rises in steps and, with a dense scatter, looks smooth. Another model may grow smoothly in each individual sample and have a similar average. Keeping only the average curve makes these explanations difficult to distinguish. Complete trajectories, distribution of transition times, or additional maintenance checks are needed.

What has now been established

We distinguished between running, saving and logging the result. These roles need to be tested with suitable interventions. Now we will see them on a small predetermined mechanism.

Practice 17 Checking startup and saving

Why do it: Compare starting and maintaining activity according to the rules of the chapter's educational model.

What you will need: State change rule and threshold; distinction between preconditions, process and consequences. Temporal proximity does not by itself establish a causal role.

Time reference: 25–35 minutes. Can be done intermittently; This is not a speed standard.

1. Start with activity 0. At each step, keep half of the previous value; input adds 0.4; if the previous activity is not less than 0.5, add 0.3. Limit the total to one.
2. Apply input at steps 1 and 2, then remove it. Predict the trajectory up to step 4 in advance.
3. Repeat without the support increase 0.3, and then without the first entry. Compare each trajectory with the full version.
4. Separately put a mark for recording the result after step 4. Check to see if this label changes the dynamics itself.

What to keep: Three trajectories, distinguishing between initiation and maintenance, and clarification of the role of subsequent recording.

Hint: The additional increment condition is checked against the previous state, and not against the already updated value of the same step.

Exploratory follow-up: Create several samples with different start times and compare their average to the smooth variation model. Suggest a metric that distinguishes between these cases.

Analysis and answers - Appendix F, chapter 17, p. 288

Self-test questions

1. Why is dynamic emergence not the same as evolutionary origin?
2. Does sustained activity prove the presence of a conscious episode?
3. What does the discrepancy between neural and behavioral indicators mean?

The dynamics of a single episode did not answer how a person’s abilities change with age. The next chapter will maintain this distinction and examine the development of several capabilities separately.

Chapter 18

How self-awareness develops

The episode occurs in a short time, and the ability to recognize oneself or ask for help develops on a different time scale. Therefore, the question of the age of emergence of consciousness must be divided into specific achievements. We will look at sensory content, visual self-recognition, uncertainty management, and biographical self-description. This separation will prevent the failure of one test from being considered the failure of all other capabilities.

Development of multiple capabilities

The question “when does a child become aware of himself” seems like a request to name his age. But first you need to decide what kind of ability we are talking about. Distinguishing between one's own movement and an external event, recognizing one's body, appreciating one's mistake, using the word "I" and reflecting on one's own reasons are different achievements. They may have common premises, but one indicator does not automatically measure all the others.

In this chapter, development means the change in the organization and capabilities of one person over time. It involves the acquisition of new ways of interacting, changes in memory, learning and control. We do not presuppose in advance a single point up to which there is no experience, and after which all forms of self-consciousness simultaneously appear. The goal of the study is to determine what exactly is changing and what observations make it possible to distinguish it.



Figure 38. Development in everyday interactions. Object illustration of early interaction with an object. Directed movement, discernment of the consequences of an action, and subsequent self-narration require different observations; the image does not set the age threshold for the appearance of consciousness.

Experience and ability to communicate

Kouider and colleagues examined brain responses to masked faces in infants of five, twelve, and fifteen months. In all age groups, a late nonlinear response was found, which was interpreted by analogy with the indicator of conscious access in adults; among the younger ones it was weaker and slower. The work provides a neural argument for an early capacity for conscious perception, but does not establish the age of the very first experience or measure extensive reflection. [28]

The logic of such a transfer requires careful attention. If an indicator is associated with experience in adults, its presence in another group supports some commonality of mechanisms. But the researcher must check whether the signal performs the same function. The absence of an adult pattern is also ambiguous: the system may solve a similar problem in a different way, or the measurement may not be sensitive enough to its organization.

This applies to more than neural data. A verbal question adds requirements for understanding words, retaining instructions, and choosing an answer. A child may fail a verbal task for reasons unrelated to the absence of the state under investigation. Conversely, a correct answer learned as a stock phrase can overstate

the sophistication of the child's understanding. It is therefore useful to compare several tasks with different additional requirements.

What does recognition in the mirror show?

Beulah Amsterdam's classic study observed age differences in children's responses to reflection; By the end of the second year, some children showed signs of self-recognition. Even within the group studied, the results were not the same. These data relate to a specific situation of visual self-recognition, and not to a universal calendar of the emergence of consciousness. [29]

Let's break down the problem into our own training scheme. You need to detect an unusual reflection detail, associate the image with the body, find the corresponding place on the body and perform the action. Failure can occur at any level. For example, it is difficult for a person to move his hand, a detail does not seem interesting, or the image is perceived in an unusual way. In order to interpret the result, control observations about each additional requirement are required.

Success also allows for several levels of explanation. It can rely on a strong connection between movement and image. A stronger statement concerns presenting oneself as the same person outside of the current interaction. Even more powerful is the ability to reason about oneself as an object of someone else’s assessment. Moving on to each next statement requires additional tasks, not expanding the meaning of one successful movement.

Let's imagine an adult program that detects a mark on the camera image and touches the corresponding part of the body. It can be specially taught such a connection. This engineering example itself shows why the behavioral criterion needs to be decomposed into mechanisms. It does not prove that the child performs the task in the same way, nor does it negate the importance of self-recognition in developmental research.

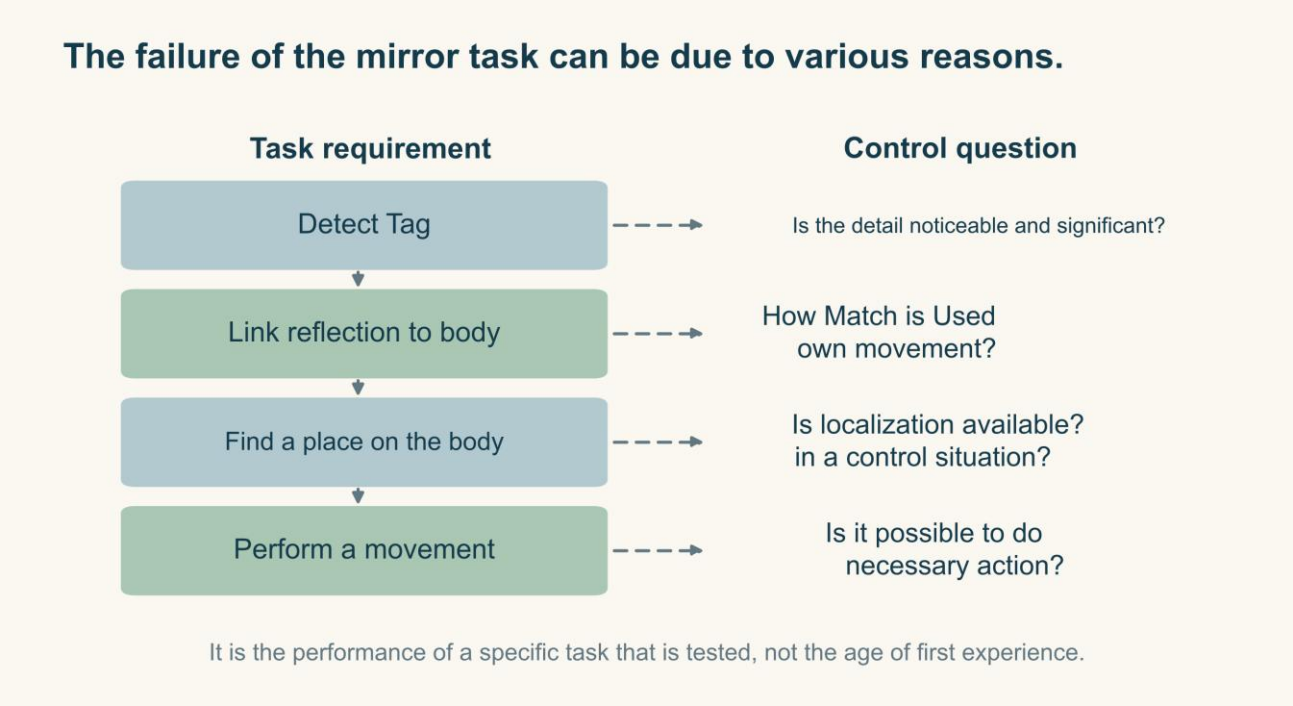

Figure 39. Decomposition of the visual self-recognition task. Touching a mark depends on several conditions: its detection, the connection of the reflection with the body, localization and the possibility of movement. Control questions help to isolate these reasons for failure, but do not themselves constitute a complete experimental protocol. One unsuccessful test cannot establish the absence of experience or other forms of self-reference.

Monitoring What One Does Not Know

In a study by Goupil, Roman Monnier, and Kuidera, twenty-month-old children were given the opportunity to request help on a memory task. Using assistance was subject to uncertainty and improved performance compared to always having to choose on one's own. The authors interpreted this as early nonverbal control of one's own knowledge. The result does not imply that the child has an adult concept of knowledge or is able to explicitly explain the reasons for the decision. [30]

To understand the mechanism, we distinguish between an assessment of the difficulty of the situation and an assessment of the quality of one’s own memory. In the first case, the child can learn the rule: after a long delay, ask for help. In the second, he uses information about how reliably he remembers a specific event. In a typical task, both strategies can coincide. A discriminative test must break this connection, for example comparing different opportunities to observe an event at the same delay.

The social side of the request is also important. For help to become a measurable action, you need to understand that the other person is able and willing to provide it. Therefore, rare calls may reflect the structure of the communication situation,

and not just confidence. Such an analysis does not invalidate behavior, but rather shows which functions work together to make it possible.

From events to biographical self

Let's look at our own example of acquiring self-description. The student makes mistakes in mental calculation several times and concludes: “I don’t count well.” This phrase combines memory of events, generalization of results and application of characteristics to oneself. If the student then discovers that errors occurred only when in a hurry, the self-description may change: “When time is limited, I choose too early.” This is a more accurate model of your own work conditions.

This transition shows the research meaning of reflection. The object of change becomes the rule by which a person explains his results. One can study whether it distinguishes between permanent ability and temporary conditions, whether it seeks additional observation, or whether it transfers the refined rule to a new task. Age alone does not explain this: information about previous experience and available testing methods is needed.

Developmental research must also distinguish between comparisons of age groups and observations of the same group over time. The first shows the differences between the people examined. The second allows you to track individual changes, but repetition of the task itself can influence the result. Both strategies require control over how comparable the measurements are.

Let’s imagine that during a repeated examination the child is better at asking for help, but the rule for asking for help was specifically explained between meetings. Improvement confirms the change in execution; it does not by itself allow for the separation of rule learning and change in uncertainty control. This separation requires a new task with a different query method. If transference persists, the narrow explanation of a single learned gesture becomes less sufficient.

Such an analysis helps to avoid the opposite conclusion. Failure of transference may reflect a lack of understanding of the new communication situation. Therefore, it is useful to separately check whether a new response form is available before considering the changed result to be an indicator of the development or absence of an internal ability.

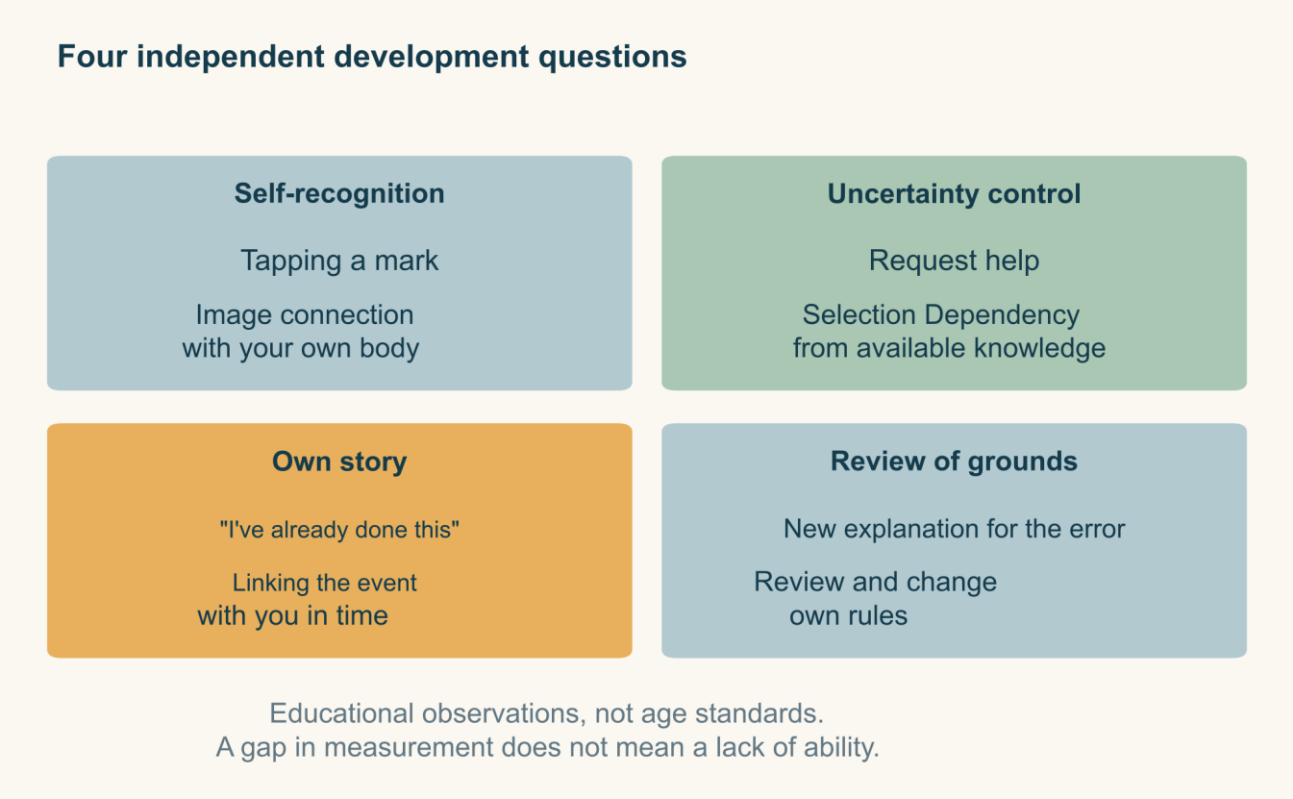

Figure 40. Development profile without a common ladder. The chapter's four fictitious case studies address different questions: self-recognition, managing uncertainty, relating events to the self, and reconsidering grounds. They should be measured separately. There are no age boundaries or a single scale of consciousness in the diagram.

How to decompose a mirror problem

To touch a mark detected through a reflection, you need to notice the mark itself, associate the reflection with the body, localize the corresponding location and perform the movement. Failure of any link can give the same external result. Therefore, the researcher first tests the availability of these activities in a control condition. Success also does not automatically transfer to adult reflection: it refers to a certain organization of visual self-recognition. Decomposing the problem does not invalidate the test; it makes the meaning of the result more precise.

What has now been established

A single calendar boundary does not replace a capability profile. One can examine the development of each relationship and then find out what background they have in common.

Practice 18 Four observations and some explanations

Why do it: Learn to distinguish between ability data and additional requirements of a specific task.

What you will need: Perspective, affiliation and authorship from chapter 3; the difference between success and self-control from chapter 4. Development here refers to changes in one organism over time.

Time reference: 25–35 minutes. Can be done intermittently; This is not a speed standard.

1. Create four lines: visual discrimination, recognition in the mirror, asking for help when uncertain, telling about your past.
2. For each, indicate the observed action and at least one additional requirement: speech, memory of instructions, movement, or understanding of the possibility of assistance.
3. For an unsuccessful result, offer two explanations and a control that can separate them.
4. Select one published chapter work and limit the summary to its task and age group, without assigning an age of first experience.

What to keep: A profile of four questions and one disassembled control of an alternative explanation.

Hint: There is no need to do home tests with your child for discussion. Use published description and case studies.

Exploratory follow-up: Suggest a transfer of ability to another form of response. First, separately make sure that the new form of action is clear to the participant, and then discuss the transfer of the internal mechanism.

Analysis and answers - Appendix F, chapter 18, p. 288

Self-test questions

1. Why can’t the age of successful self-recognition be considered the age of the emergence of consciousness?
2. How does asking for help differ from explicit reflection?
3. What does the absence of the expected indicator for the younger group mean?

The development of an individual has its own data and limitations. Now let's move on to another story: the origin of mechanisms in evolutionary lines.

Chapter 19

How to Explore the Evolution of Experience

The development of one individual and the history of the origin of the species require different evidence. Let us now consider how comparisons of function, design, and relatedness limit the story of the evolution of mechanisms. The usefulness of memory depends on the environment, and a similar ability may have a different history. Therefore, we will not build a single ladder of consciousness. The purpose of the chapter is to explicitly examine the transitions from observed feature to historical hypothesis and from mechanism to putative experience.

History of the mechanism and history of the experience

The structure of a living organism has a history of origin. It is therefore natural to ask when the experience-related mechanisms arose and what tasks contributed to their maintenance. But reconstructing the history of a mechanism and establishing the presence of an experience are different research operations. One can explain well the benefits of memory and still argue about the organization of memory that is accompanied by conscious recollection.

The evolutionary question must be formulated narrowly. For example: under what conditions could selection support integrating multiple sources of information, controlling uncertainty, or distinguishing between one's own actions and external events? Such formulations specify functions that can be modeled and compared. The next question—whether these functions are conditions of experience—requires a separate theoretical connection. Confusing questions creates overly confident narratives about the origins of qualia.



Figure 41. The environment sets different behavioral tasks. Subject illustration of different living conditions near water and on land. The benefits of memory, retrieval, and action selection depend on the specific environment; The arrangement of organisms in the figure does not determine the order in which the experience occurs.

How explanation through advantage works

Let's consider our own model of two inherited strategies. The first one chooses an action every time based on the last signal. The second saves several recent observations. The world sometimes changes, and the sensor makes mistakes. In a stable world, memory can reduce the impact of random errors. In a rapidly changing world, the same memory can delay switching and make behavior worse. Therefore, the advantage of a mechanism depends on the environment and its parameters.

We will add the cost of storage and processing. If winnings are rare, even an accurate memory may not compensate for the costs. Selection in this training model depends on how we have specified the relationship between successful actions, costs and reproduction. Therefore, the phrases “more difficult means better” or “more memory means closer to consciousness” do not follow from the constructed task. We explore trade-offs between features rather than movement towards a pre-assigned apex.

Now let's supply the second strategy with a variable called "experience." If a variable does not participate in any way in the causal organization, adding it will

not change anything in the selection model. This shows the limitation of a particular design. It does not follow from this that experience in nature is causally powerless. Perhaps it is identical to some active process; perhaps the supposed process connection requires a different explanation. The model must explicitly indicate the selected option.

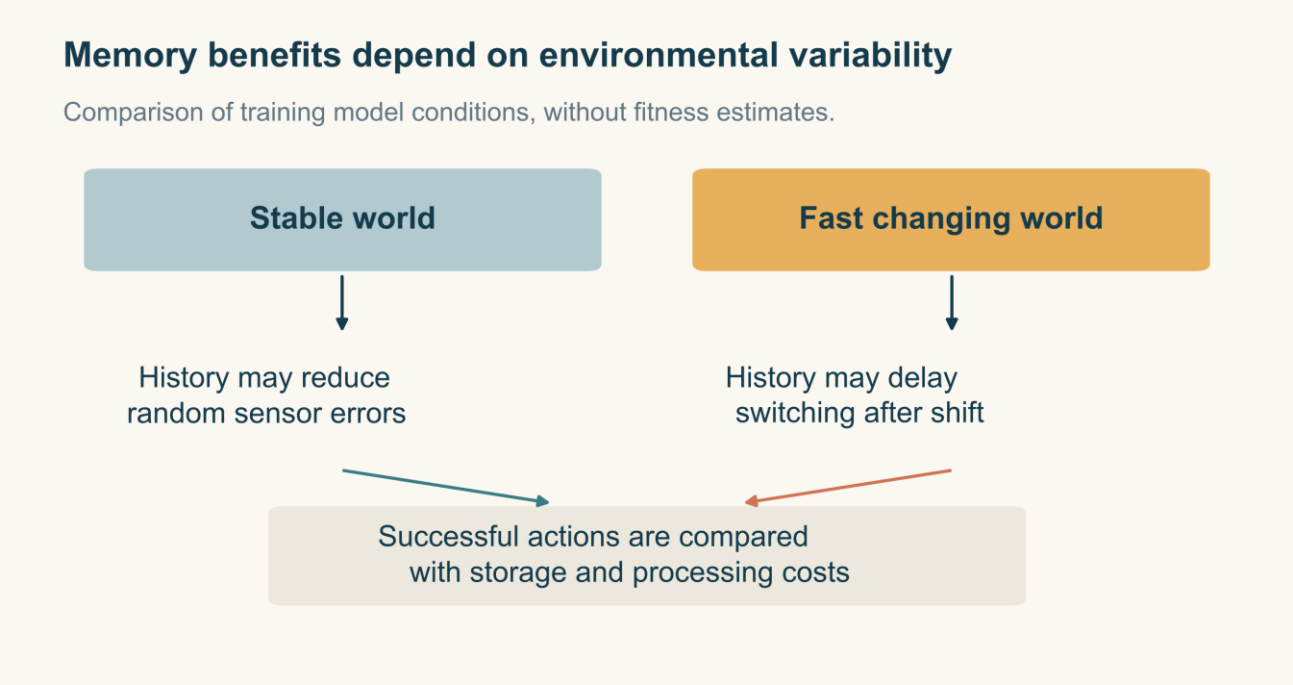

Figure 42. Memory trade-off instead of complexity ladder. In the chapter model, memory is able to smooth out noise when the target is stable and retain outdated information after it has changed. To draw a conclusion about an advantage, success and costs must be taken into account independently, and the connection with reproduction must be explicitly stated. This comparison of functions does not establish an evolutionary origin for the experience.

Similarities do not tell one story

If two species are found to have a similar ability, different historical explanations are possible. It could come from a common ancestor, be formed independently, or be based on a combination of common and new elements. In addition, the same result of a task can be achieved by different mechanisms. Therefore, comparison requires a joint analysis of behavior, design, and origin, rather than counting similar responses.

Felsenstein showed why species cannot be statistically treated as completely independent observations: their characters are linked by a common history of origin. His methodological work suggests taking kinship into account in comparative analyses. For the study of consciousness, this principle means that ten closely related species with one trait are not equivalent to ten independent cases of its occurrence. [31]

Let's build a fictional tree with four finite branches. Two close branches have the M mechanism, the other two do not demonstrate it. One explanation suggests the emergence of M in the common ancestor of the first pair. The other is an earlier appearance followed by loss in the second pair. Based on the distribution of one attribute, the story may remain ambiguous. Additional information can distinguish between versions, but omissions in observations cannot automatically be considered a loss.

When an experience is studied, uncertainty increases. Even the presence of M is not always established with equal reliability, and the connection of M with experience may depend on the theory. Therefore, historical reconstruction must preserve both uncertainties. It is incorrect to first declare a controversial indicator to be an exact criterion of consciousness, and then present the confidence of the constructed tree as an independent confirmation of the criterion.

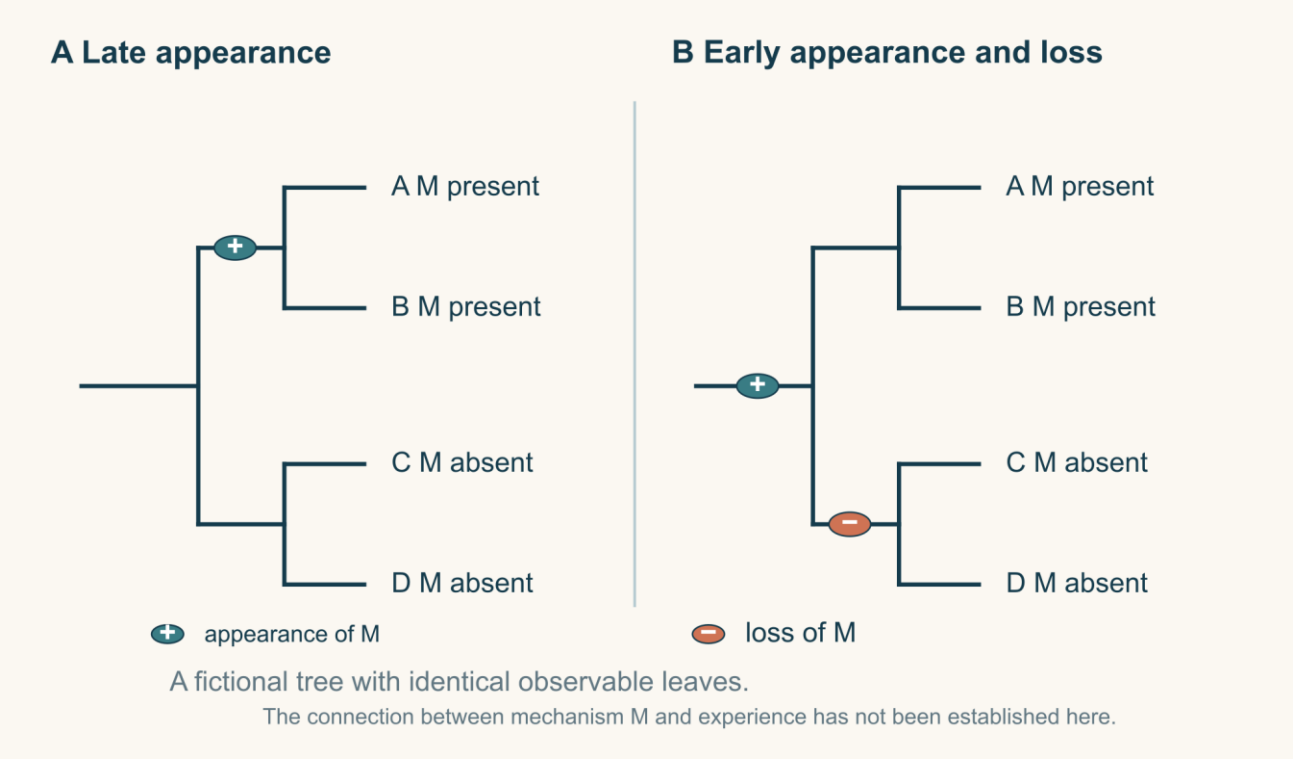

Figure 43. Two histories of the same trait distribution. In the teaching tree, mechanism M is present in species A and B and absent in C and D. This is compatible either with M first appearing in the common ancestor of the first pair, or with M arising earlier and later being lost in the second pair. The tree does not reconstruct the origin of consciousness; the relationship between M and experience remains a separate question.

Comparing Function and Organization

Barron and Klein proposed an argument about the possible experience of insects that relies on functional analogies between the behavioral control systems of insects and vertebrates. This is a primary theoretical work, not a direct observation

of qualia and does not date their origin. Its research value lies in the explicit connection between the assumed organization and the inference about the experience; the transition itself remains subject to scrutiny and debate. [32]

You can start your own comparative project with a much more modest question: does the system integrate information about space, available actions, and internal state when making choices? To answer, you will need to show that changes in these factors interact and do not cause three independent reactions. For example, the availability of a resource should change the choice of route depending on the cost of movement, and new information should be transferred to a previously unseen combination of conditions.

Such a result would support the hypothesis of a specific management organization. Causal mechanisms should then be sought and simpler explanations tested. The transition to consciousness will depend on why this particular organization is considered essential. This is where the material in chapters 13–16 becomes useful: different theories guide comparative research towards different attributes and interventions.

Why does a universal ladder get in the way?

One scale forces disparate abilities to be combined. A species may be good at a spatial task but poor at a task that requires an unfamiliar sensory channel. Another system may accurately estimate the uncertainty of one action, but not transfer the estimate to a new area. These profiles do not have to be ordered in one line.

A map of functions is more useful for analysis: available types of information, time horizon, transfer flexibility, bodily organization, error control. If phenomenal experience is of interest, a separate column of arguments about it is placed next to it. If explicit reflection is of interest, another set of tasks is required. This idea prevents the subtle conclusion that “better in one test means more conscientious in everything.”

One should also not turn the modern look into a living step of the past. It has its own history of changes. Comparison helps to reconstruct certain ancestral traits, but does not make the observed animal an exact copy of a distant ancestral state. Therefore, one's own origin scenarios need to be presented as models with explicitly stated assumptions.

It is useful to check the script for a circular explanation. Let's imagine the reasoning: the mechanism was preserved because it was useful; its usefulness is established by its survival. Here one statement supports itself. To break the cycle, one must independently measure the perceived benefit, the cost, and the

conditions under which the benefit disappears. Then the model may be wrong in terms of content.

In the memory example, the condition being tested would be a sudden change in the environment. If the model predicts a loss of benefit, but it persists across all parameters, the implementation and costing should be examined. Perhaps the comparison is designed so that one strategy always receives extraneous information. A technical error like this can produce a compelling but false adaptation story.

Identical tree leaves allow two stories

On the training tree, species A and B found mechanism M, while C and D did not. One story suggests the appearance of M in the common ancestor of A and B. Another allows for an earlier appearance and subsequent loss in another branch. The observed ends of the tree are the same, the historical changes vary. To make a choice, you need additional information about relatives, device and likely transitions. If the connection of M with experience itself remains a hypothesis, confidence in historical reconstruction does not turn it into independent evidence of consciousness.

What has now been established

The history of the mechanism and the history of the experience did not merge into one conclusion. It is possible to explain the advantages of a specific memory and at the same time keep open the question of in what organization it has an experienced side.

Practice 19 Map of the Evolutionary Argument

Why do it: Construct an evolutionary argument with explicit assumptions at each transition.

What you will need: Distinguish between development, evolutionary origin and the presence of experience. The mechanism, the observed trait, and the putative ancestor are different elements of the argument.

Time reference: 25–35 minutes. Can be done intermittently; This is not a speed standard.

1. Use the tutorial drawing tree 43. Write your observations on leaves separate from the two versions of the story.
2. For each option, name the expected appearances and losses of the mechanism. Do not pass off any variant as a direct observation of an ancestor.

3. Describe an environment in which multiple-signal memory can help and an environment where it delays switching. Please also include storage and handling costs.
4. Add a separate sentence about the relationship between mechanism and experience and label it as a theoretical hypothesis. Provide information that would change each conclusion.

What to keep: Map from observation, historical explanation, advantage model, and single phenomenal proposition.

Hint: Similar behavior does not guarantee a common device. A common molecule does not guarantee the same function of the entire system.

Research Continuation: Select a published comparative argument for the chapter. Check what information relates to function, structure, and relationship, and offer an observation that distinguishes nearby historical versions.

Analysis and answers - Appendix F, chapter 19, p. 289

Self-test questions

1. Does the adaptive benefit of a mechanism prove the presence of experience?
2. Why don't closely related species provide completely independent evidence?
3. Why save multiple origin stories?

We separated episode dynamics, development and evolution. The next chapter will highlight the remaining question: why a certain physical organization can have an experiential side at all.

Chapter 20

What an Explanation of Emergence Requires

Previous chapters explained how a process emerges over time and how its capabilities change. Now let us ask about the basis of the very presence of experience. A successful vision model can predict responses and errors, but its author must be explicit about what connection between computation and experience he is proposing. This requirement sets up the book's own hypothesis. It allows for strong assumptions to be made while preserving their commitments and possible objections.

Clarification of the remaining question

Let's assume we've built a detailed vision model. It explains recognition, memory, confidence, and response preparation. When changing parameters, errors similar to human ones occur. What else is required to call it an explanation of experience? The answer depends on what connection between the physical process and experience we accept and what data we justify this connection.

David Chalmers has formulated a distinction between explaining cognitive functions and asking why the performance of functions is accompanied by subjective experience. This argument poses a philosophical problem, not an experimental proof of a separate non-physical entity. One can not accept his further conclusions and still use the requirement to accurately designate the transition from mechanism to experience. [2]

In this chapter we are interested in the ontological meaning of emergence: why some process has an experiential side at all. Information about when the network turned on, when the child mastered the task, or when the mechanism appeared in the history of the species does not automatically answer this question. However, they limit the possible answers and help turn general ideas into research programs.

Explain the presence and explain the difference

Let's start with the difference between the two tasks. First: why is experience present in one state of the system and absent in another? Second: why does present experience have this particular quality? An explanation that predicts only the general possibility of consciousness does not yet explain the difference between sound, color and bodily sensation. Conversely, a color discrimination

model can accurately describe the relationships between responses while leaving the question of experientiality open.

For our own example, let’s take a space of three sensations, designated by conventional letters. The participant considers the first to be closer to the second than to the third. The model reproduces this similarity structure. This is a meaningful achievement: it explains certain relationships in the data. But the coincidence of the three distances does not establish that the internal states of the model have the same qualities. It is necessary to understand what properties of space are being measured and why the chosen correspondence has explanatory power.

The next step is to predict changes. If an intervention changes one connection in a mechanism, should the similarity of the two sensations, their unification, or the accessibility of discrimination change? A prediction made before a measurement is stronger than simply choosing a pretty picture after the result. Selectivity is also important: changing one relationship should preserve others if the model asserts their independence.

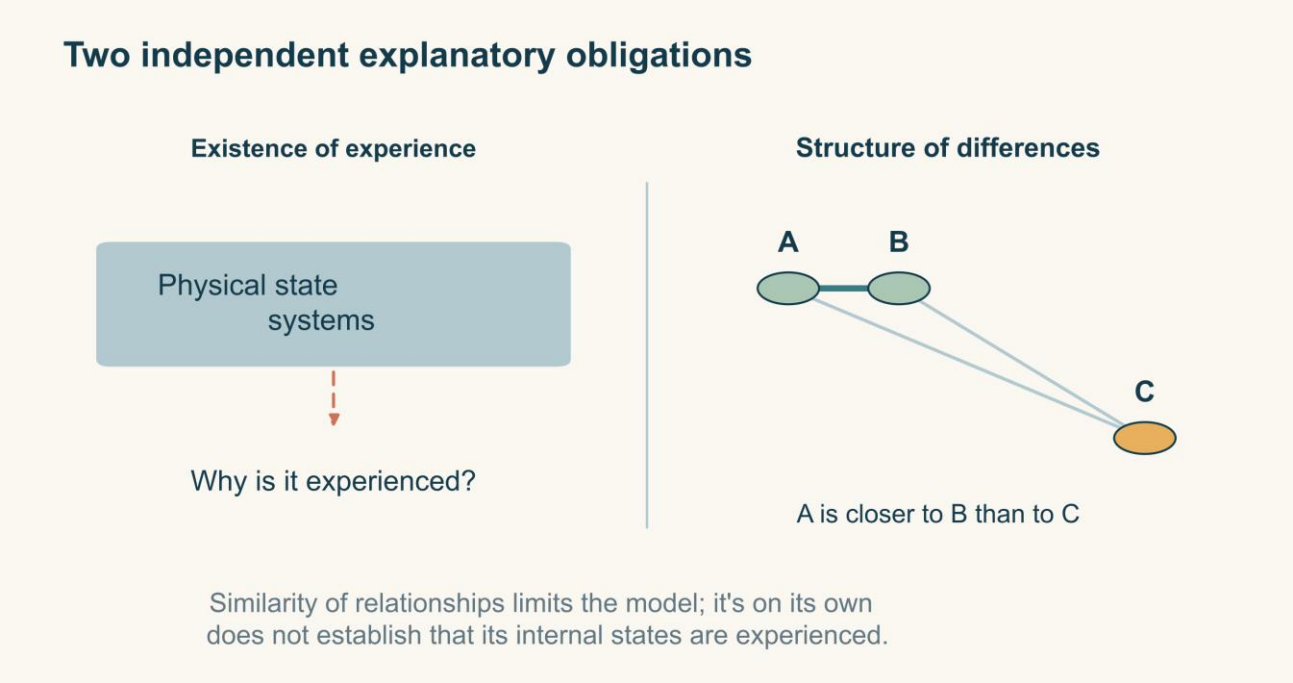

Figure 44. Explaining presence and explaining quality. On the left is the question about the presence of experience, on the right - about the structure of the differences between the three conventionally designated sensations. The geometry on the right shows only a given proximity relationship, without numerical distances or reference to modality. Reproducing such a relationship is a limitation for the model, but does not automatically solve the left problem.

Unity and belonging

Sound, visible movement, and bodily sensation may all be present together in one episode. The explanation should indicate why they are included in the same

episode, as well as why some system processes are not included in it. A simple sum of records does not explain this: one must define the relationships through which changes in one part affect the rest and the accessible whole.

The identity of the experience also requires clarification. In a minimal sense, we are talking about the perspective in relation to which the content is organized. In a stronger sense, a person recognizes the state as his own. An even stronger version includes the concept of oneself as a participant in events that persists over time. If a theory assumes that an explicit concept of self is necessary for all sensation, it assumes an additional commitment that cannot be hidden within the definition of consciousness.

Let's check this with a mental example. The person is absorbed in the melody and does not analyze himself. Later he may describe the experience in the first person. Was the extended self-description part of the original episode or did it appear during the report? The story itself does not resolve the issue without additional clarification. But the example shows why the perceptual model, the belonging model, and the reflective model are useful to first distinguish and then explore their interaction.

What relationships are possible between process and experience?

One version of the explanation assumes identity: experience is a certain physical process described from the other side. Then it is necessary to justify the choice of process and show why its structure corresponds to the structure of experience. Simply changing the name is not enough. The success of such an explanation would depend on a consistent system of predictions, causal evidence, and the absence of arbitrary exceptions.

Another option introduces an additional principle of correspondence between physical organization and experience. Then you need to explicitly write down the principle and explain what cases it distinguishes. If it does not predict anything beyond already known behavior, the question arises of how to test it and prefer it to the competitive principle. This is not a ban on research, but a requirement to show the way to obtain new information.

The third option revises the original account of phenomenality: perhaps part of the apparent properties of experience are related to the way the system represents its own workings. Then it is not enough to declare difficulty an illusion. It is necessary to explain specific judgments about experience, their stability, errors and connection with other data. Here again, a strong answer should produce testable differences, rather than end up renaming the question.

These options are given as logical strategies. We do not consider them equally confirmed and do not choose a winner based on one mental case. It is more useful for a research textbook to learn to determine exactly what step each model takes and how much explanatory debt it has left.

How to formulate your own hypothesis

Consider a limited research hypothesis: the stable integration of visual and bodily content depends on whether the system derives them from a single, modifiable common cause. The hypothesis concerns a certain organization of the episode. It predicts that the violation of precisely this unification will change the joint belonging of contents while maintaining their separate distinction. This is our instructional formulation, not a statement of established universal law.

Testing will require identifying common cause in the model, finding independent evidence of separate discrimination, and selecting an intervention. Then compare the explanation with an alternative that attributes the outcome only to attention or memory. If all models produce the same expectations, the conditions should be changed. The mere presence of the words “common cause” does not make the hypothesis deeper than the existing descriptions.

Even a positive result would be supported primarily by the unification mechanism. To call it an explanation of the presence of experience, one must add and justify the corresponding principle. A possible scientific work therefore consists of several related achievements. To require that each small experiment solve the entire ontological question at once would be to lose useful intermediate constraints. Passing off restrictions as a final solution would be another mistake.

A good explanation should also indicate the scope of applicability. The visual unification hypothesis may say nothing about dreaming, bodily pain, or abstract thinking. This is acceptable if the restriction is written down in advance. Universality is not achieved by simply adding the word “every” to a successfully verified particular statement.

Expanding the field requires new connecting assumptions. It can be suggested that different types of content use the same organizing principle, but implement it with different elements. Then it is necessary to determine what exactly remains common and what change in each area is considered similar. If a match is chosen arbitrarily after the results, the program loses verifiability. If it is given in advance, a path arises from a particular mechanism to a more general theory of experience.

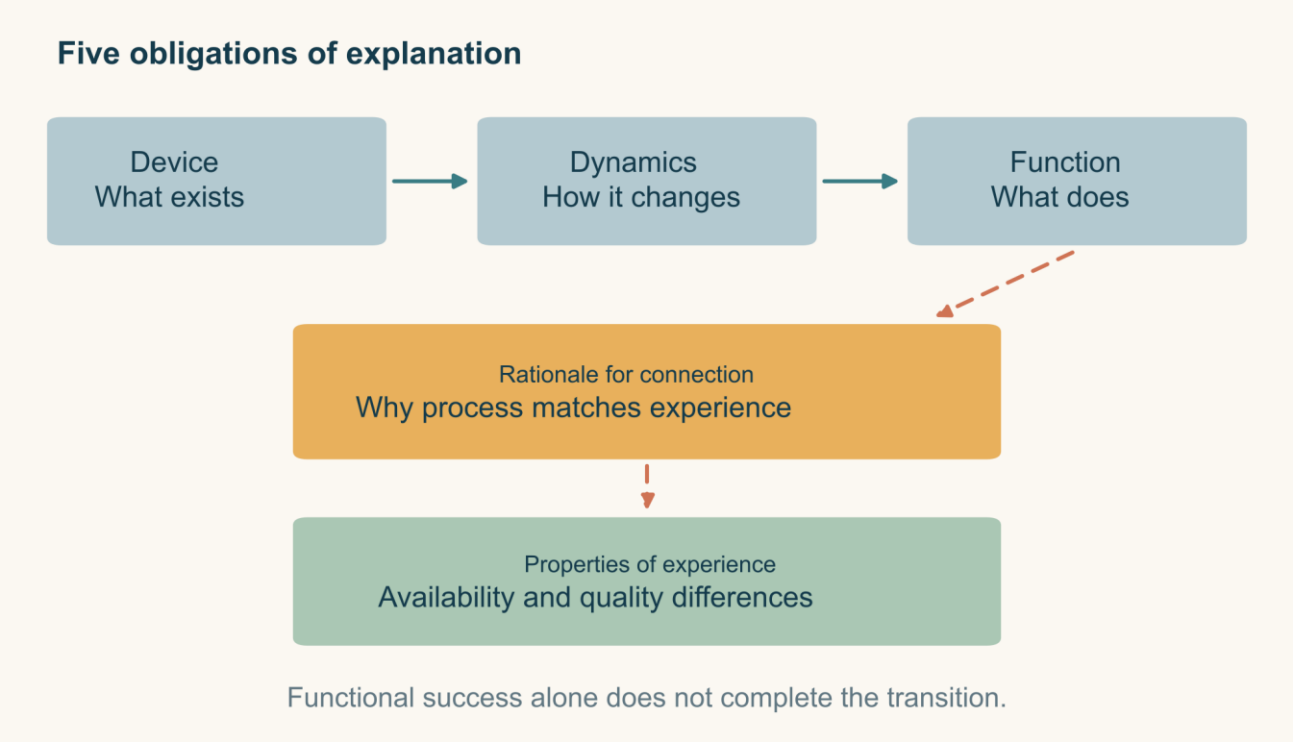

Figure 45. From device to explanation of experience. Five questions from the explanation audit: what exists, how it changes, what function it serves, what properties of experience are being explained, and why the process is associated with them. The dotted line marks a transition that requires its own justification. An explanation of a function does not automatically establish the presence of experience or its qualitative structure.

Five sentences do different explanatory work

The description of the associated elements answers what exists in the device. The rule for updating them shows the dynamics. Successful discrimination of objects describes a function. The relationship between the communicated qualities specifies a phenomenal requirement. The principle of relevance explains why the chosen organization is considered sufficient or essential for the experience. If we leave only the first three points and call the exit an experience, the last transition will not disappear. If we propose an explicit identity of process and experience, it receives substantive commitments: a structure of correspondences, boundaries of application and discriminating consequences.

What has now been established

We have identified what explanation is still missing after a good functional model. Now let's try to formulate this remainder precisely, without abandoning the result already obtained.

Practice 20 Audit explanations

Why do it: Break down one explanation for its individual obligations and find a strong competitor.

What you will need: Necessity, sufficiency, correlation and presumptive identity from chapter 5. Distinguish between the presence of experience, its quality, unity and belonging.

Time reference: 30–45 minutes. Can be done intermittently; This is not a speed standard.

1. Take the specific explanation from chapters 13–15. Write one sentence each about the structure, dynamics, function, explained property of the experience, and the connection between them.
2. Next, indicate the basis for each sentence: definition, data, calculation, or theoretical assumption.
3. Select a competitor that retains some of the same functional results. State what additional requirement distinguishes the explanations.
4. State a possible discriminating observation. If it is not yet defined, describe exactly what method or output is still required.

What to keep: Five related statements, a competing explanation, and a precise remaining question.

Hint: Explaining the relationships between qualities is a meaningful result, even if the presence of experience in the new realization has not yet been established.

Exploratory follow-up: Propose your own limited correspondence principle. Test it on a simple controller and on a system with the same responses received by another device.

Analysis and answers - Appendix F, chapter 20, p. 289

Self-test questions

1. Why might an explanation of color discrimination fail to address the question of experience?
2. Is it necessary to introduce an additional entity to explain the experience?
3. What is considered progress if there is no definitive explanation yet?

Before choosing a device for a future candidate, let's consider data on different media of experience. This will show why the criteria need to be applied to specific mechanisms and conditions.

Part 6

Possible bearers of experience

Chapter 21

Consciousness and self-control in animals

The theoretical principle must retain its meaning during the transition from a familiar person to another organism. However, the literal repetition of a human task may require inappropriate speech, movement, or sensory channel. Consider animal studies as distinct arguments about memory, content, and post-injury states. Their strength lies in the combination of appropriate measurements and control of alternatives, not in the superficial resemblance of the answer to a human phrase.

What questions do we ask the animal?

The question of an animal's possible experience requires distinguishing between several purposes. Phenomenal consciousness means having an experience. Self-reference concerns the connection of information with one's own body, action, or state. Metacognitive control uses information about the reliability of cognitive processes. Explicit reflection makes one's own reasons the subject of targeted consideration. These capabilities cannot be considered interchangeable, even if they interact in some tasks.

An animal that avoids uncertain choice may evaluate the quality of memory, use a learned external cue, or follow a simpler rule of reinforcement. An animal protecting an injured body part exhibits a different set of functions. No observation should automatically translate into a statement about the human way of thinking about itself. We need a profile of specific reasons and open alternatives.



Figure 46. Different organisms and different research questions. Subject illustration: bird, octopus, fish and bee. The scale is relative. Possible bearers of experience should be compared based on specific characteristics and research; The position and size of the animal on this page does not define a scale of consciousness.

General assessment and specific experiment

New York Declaration 2024 reports strong scientific grounds for attributing experience to other mammals and birds and the realistic possibility of experience in all vertebrates and many invertebrates. This is a statement from the signatories, not a separate experiment and not the same evidence base for each species. It is useful to use as a guide to research uncertainty while maintaining differences between specific data. [33]

Overall assessment can draw on multiple lines of inquiry, but individual learning assessments must remain testable. Therefore, below are several primary works on different subjects: memory, visual content, and damage-related states. They do not constitute an animal competition for a single point of consciousness. Each work limits its range of explanations.

Memory and visual content

In Hampton's study, two macaque monkeys could accept or reject a memory test. Choice depended on memory conditions and was associated with subsequent accuracy; The results of the animals differed in the completeness of fulfillment of the criteria. The work gave grounds to discuss control over the availability of one's

own memory. It did not show verbal reflection and did not immediately exclude all possible associative explanations for any such problems. [34]

In a study by Nieder, Wagener and Rinnert, the activity of crow forebrain neurons was related to the behavioral report of the presence or absence of a weak visual stimulus. The authors interpreted the result as a neural correlate of sensory awareness. The study expands the comparative base, but does not define all the properties of avian experience or establish the presence of human biographical self-awareness. [35]

The two works answer different questions. The first helps to explore decision control based on available knowledge. The second links neural activity to discrimination of reported visual content. Shared consideration is useful precisely because of difference. Calling both results simply “self-awareness” eliminates the possibility of understanding which mechanism explains each effect.

Damage and negative experience

Crook's work with octopuses examined avoidance of a location associated with a noxious stimulus and preference for a location associated with relief after it. The author interpreted the behavioral and neurophysiological results as evidence to suggest an affective component of pain. This is an argument about the experienced negative state under the conditions studied, not a direct measurement of alien qualia, nor is it an argument for the same experience of all cephalopods. [36]

Barr and Ellwood in 2024 examined how shore crabs relate the avoidance of aversive stimuli to the avoidance of bright light. The variation in choice across competing conditions was consistent with the predictions of flexible evaluation. The authors considered the outcome to be consistent with pain. The narrower implication directly demonstrated concerns the trade-off between different bases of choice. [37]

In a published study by Manzi and colleagues 2026, house crickets groomed their antennae longer after a damaging heat treatment than after control treatments. Changing the conditions, intended to be less or more stressful, did not produce a significant effect. The work strengthens the rationale for discussing directed self-defense and possible pain-like conditions; it does not demonstrate metacognitive analysis or explicit reflection. [38]

It is important here to distinguish between damage registration, organized avoidance and negative experience. A simple defensive response provides less information than a sustained change in preferences while controlling for alternatives. But even the complexity of behavior is not an independent measure of

pain. The explanation must show why the set of observations fits better with the inferred state than with the nearest competing mechanisms.

Comparative foundation map

The table combines the considered data types. It indicates the direction of testing and the boundary of transfer, rather than assigning numerical probabilities of consciousness to species.

Group	The basis for this analysis	Significant unresolved issue
Mammals	General assessment of strong bases and experience with macaque memory	What forms of control transfer between tasks?
Birds	General assessment and neural study of crows	How generalizes the result to other contents and types?
Other vertebrates	The declaration recognizes the realistic possibility of experience	What specific mechanisms and indicators are comparable?
Cephalopods	Combination of preferences and neurophysiology in octopuses	How the quality and organization of the intended experience works
Decapod crustaceans	Trade-offs of selection among shore crabs	What alternatives to assessing the experienced state are excluded?
Insects	Directed self-defense in house crickets	How are defensive control and possible phenomenality related?

In the line of other vertebrates, a separate experiment is not analyzed here. This is the limit of the selected material, which cannot be translated into a conclusion about the lack of experience. Likewise, data from one type should not be generalized to the entire group without additional comparative justification. The volume of research and the properties of the organisms being studied are two different things.

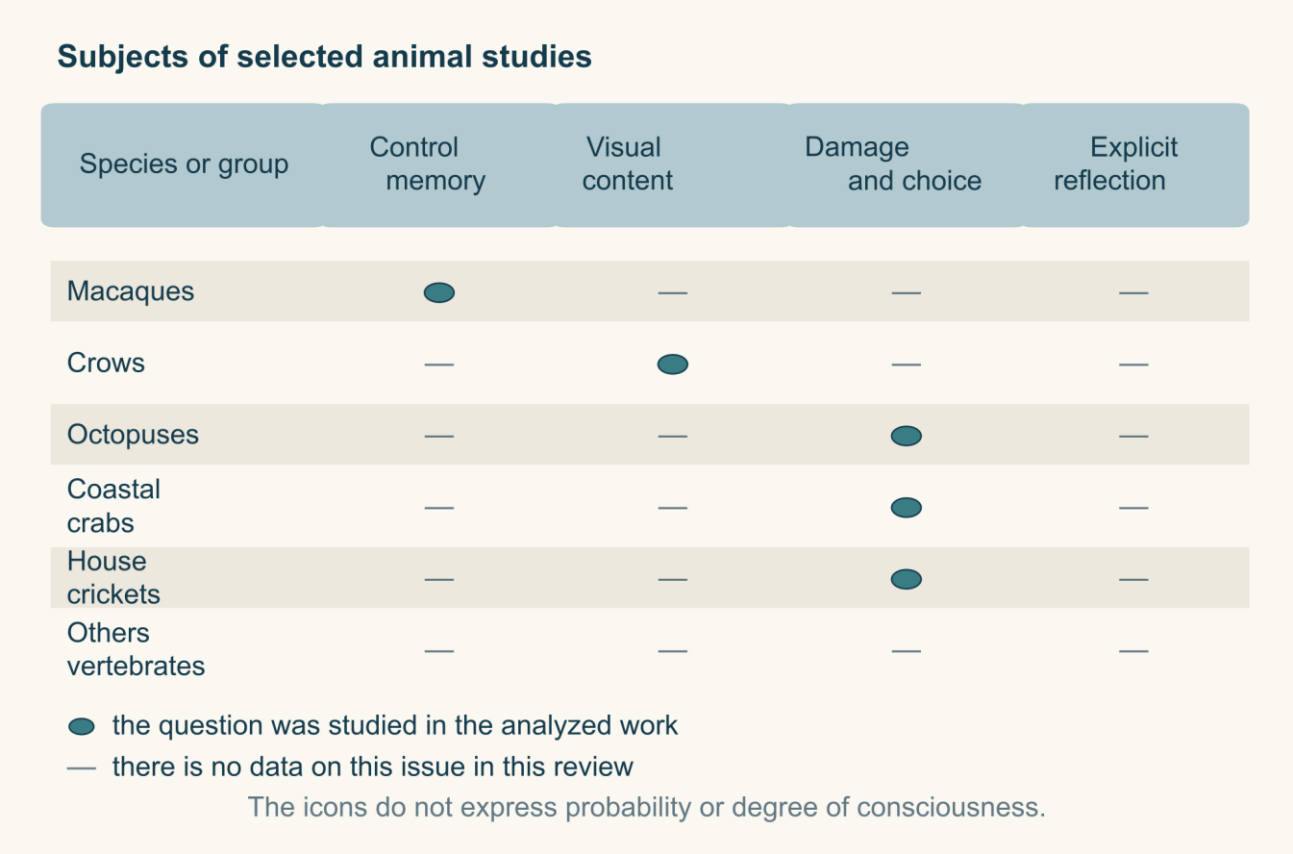

Figure 47. Matrix of questions and evidence gaps. The matrix applies only to the studies discussed in the chapter: memory in macaques, visual communication in crows, avoidance and facilitation in octopuses, choice trade-offs in crabs, and self-defense in crickets. The dot denotes the subject of the work, not the proof of the experience. A dash means a lack of appropriate analysis, and not a lack of ability in the animal. One type of data does not represent the entire group.

How to strengthen the conclusion

A strong profile is built from convergent data with various constraints. If learning transfer, sensitivity to changes in outcome value, and corresponding causal changes are observed, it becomes more difficult for a simple fixed rule to explain everything at once. However, the alternative needs to be formulated specifically. The phrase "it's just instinct" is not a mechanism unless the inputs, internal states, and rules for changing behavior are specified.

In the same way, the expression “an animal understands” requires clarification. What exactly is presented? How does knowledge manifest itself in a new situation? What changes if you remove the external cue? Can the perceived assessment of one's own error influence action before the result is obtained? The answers make the debate about explanatory richness productive.

For explicit reflection, the bar of affirmation is higher than for uncertainty control. It is necessary to show consideration of one's own reasons and their revision,

excluding the available narrower rules. The absence of such evidence does not disprove the possibility of phenomenal experience. It limits the specific attribution of reflective capacity.

When comparing, you should check whether the task itself is equally understandable. If one species is presented with a habitual choice method and another is presented with a difficult movement, the difference in accuracy will confound the ability of interest with the test design. We need to control the availability of the signal, motivation and the ability to complete the response. Sometimes a fairer comparison uses different external actions but the same relationship between information and choice.

Let's imagine two training procedures: pressing a button and selecting a passage. In both, you can evaluate how much the decision depends on the quality of the available information. This dependence should be compared, and not the external similarity of the movement. This transfer also requires justification, but it shows why literally the same test is not always the same test of the mechanism.

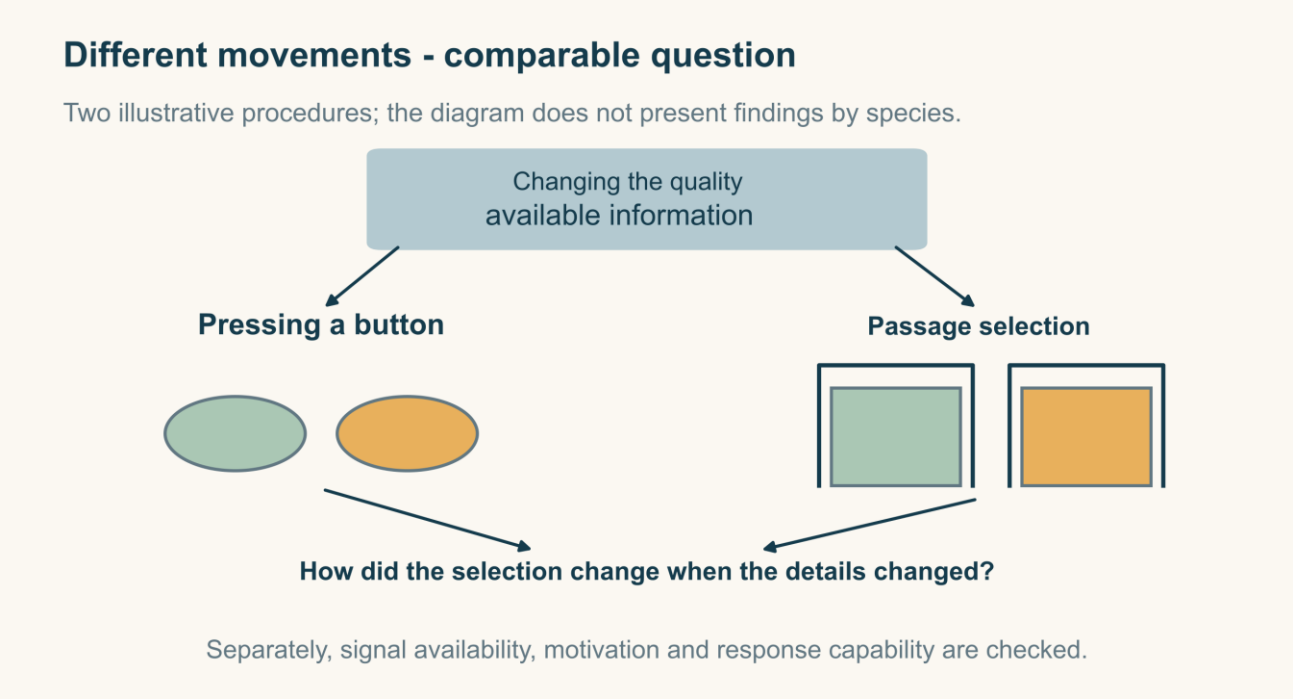

Figure 48. Comparable tasks with different external responses. Pressing a button and selecting a passage can test similar dependence of a decision on available information. When comparing, it is necessary to justify the comparability of this dependence and separately check the sensory, motivational and motor requirements. The coincidence of the external result does not make the internal mechanisms identical and does not provide a common scale of consciousness.

What does refusing a difficult test add?

The animal more often rejects the test after conditions of poor memorization, and performs the accepted tests more accurately. This is compatible with memory

availability control. A competitor may suggest a choice based on external characteristics, such as delay. To distinguish between versions, you need to change the quality of memorization at a comparable delay or introduce another suitable condition. Therefore, specific criteria and data for each animal are important. The general word “metacognition” does not replace this analysis and does not attribute verbal adult reflection to the animal.

What has now been established

Individual works answer different questions. The profile of the bases allows them to be combined while maintaining the distinction between the investigated property and the unknown.

Practice 21 Profile of one species studied

Why do it: Profile a single study without generalizing its findings to an entire group of organisms.

What you will need: Research specification, difference between functional and phenomenal inference. A declaration of scientific position and a specific experiment have different statuses.

Time reference: 30–45 minutes. Can be done intermittently; This is not a speed standard.

1. Select chapter work on memory, visual content, or post-damage status. On the first pass, fill in the fields according to the text of the book.
2. Indicate the type, task, measurement, control and main result. If the number of animals is not given in the available description, mark the blank.
3. Suggest a specific alternative mechanism and a condition under which he would expect a different outcome.
4. State one strong conclusion and one broader conclusion that this work does not yet resolve.

What to keep: One study specification and a testable alternative without an overall conscientiousness score.

Hint: The absence of studies of the selected property in this review does not constitute an observation of its absence in the body.

Research follow-up: Open the primary article and clarify the sample, exclusion rules, and individual results. Check to see if the controls you suggest are actually provided.

Analysis and answers - Appendix F, chapter 21, p. 290

Self-test questions

1. Why does recognizing possible pain not mean recognizing explicit reflection?
2. Is it possible to extend the result of one species to all its relatives?
3. What to do with an explanation through a simple rule?

Now let's check what functions are available to systems without neurons. This will help distinguish memory and adaptation broadly from the complementary experiential hypothesis.

Chapter 22

Boundaries between adaptation and experience

Complex coordination and past influences occur outside of nervous systems. To understand the significance of such data, we start with the response being measured and several possible mechanisms. The weakening of the response may arise from specific habituation or from general exhaustion. Comparison with a new signal and a pause will allow these versions to be separated. The question of experience will require the next, separately justified step.

Why systems without neurons are important

If we want to explain the conditions of experience, we cannot a priori limit our search to human-like creatures. At the same time, any useful reaction cannot be considered a sign of qualia. Studying plants, single-cells, and other nonneuronal systems is useful precisely because it forces us to separate signal processing, history storage, learning, and putative experience.

The initial question should concern the phenomenon being measured. Did the reaction change after repeated exposure? Has the change persisted? Does it depend on a specific signal? Can an organism use connections between events? Only after the answers do the following question arise: what grounds connect the found mechanism with experience? This arrangement leaves the study open-ended while protecting against over-inference.

Signaling and coordination

Toyota and colleagues showed in Arabidopsis that damage is associated with a glutamate signal, the propagation of changes in calcium concentration and the activation of defense reactions in distant parts of the plant. This is the primary study of the mechanism of long-distance communication within the body. The similarity of individual molecules and modes of signal transmission with known neural processes does not establish the identity of the overall organization or the presence of pain as an experience. [39]

Using our own example, let’s imagine a system of several tanks. A change in one triggers a signal that closes valves in the others. There is local detection, information transfer and coordinated protective action. If the signal depends on previous influences, the influence of history will also appear. All of these properties

can be explored without attributing self-beliefs or negative experiences to the system.

The analogy does not explain the whole plant. It shows a logical boundary: the presence of signal transmission does not determine what additional properties are implemented. A stronger claim would require data on the interaction of processes, choice flexibility, internal state, and the theoretical basis of the connection to experience. The word "defense" describes a function and does not have to imply a conscious intention to defend oneself.

Memory as the influence of the past

In a minimal functional sense, memory exists where history changes the current response given the same present input. Such memory may be distributed by system state and not contain a separate event record. It is necessary to distinguish between this broad meaning and memory as an experienced return to the past. The same word does not make the functions the same.

Let's build a training model. It has an habituation variable that is initially equal to zero. After each presentation of the signal, it approaches one by a fifth of the remaining distance. The response is equal to the difference between one and the current habituation. The first reactions will be one, eight tenths, sixty-four hundredths and so on. During the pause, the variable is reduced by a fifth each step, so the reaction is restored.

All numbers are given by us for the sake of simplicity. The model demonstrates how maintaining a state creates response reduction and recovery from an interruption. There are no neurons, no verbal beliefs, and no observer inside the model. But from a single reaction curve it is impossible to know whether such a rule coincides with a real biological mechanism. Selection requires additional interventions and measurements.

Let's add two signals and two separate habituation variables. Then repetition of the first signal reduces the reaction only to it, while the second remains effective. If you use a single shared resource variable instead, both responses will decrease. Thus, testing the new signal helps distinguish between specific change and general attrition. It is still not a test for phenomenal experience.

What is shown in slime mold

Boisseau, Vogel and Dussutour studied the slime mold Physarum polycephalum, an organism without neurons. With repeated exposure to chemicals, the reaction weakened, and after a break it was restored; The specificity of the change was also analyzed. The authors interpreted the results as habituation, a simple form of

learning. The study shows the possibility of such plasticity without a nervous system. It does not measure qualia, self-awareness, or overt reflection. [40]

Here it is useful to separate the recognition of the result from the expansion of its meaning. If the habituation criteria are met, there is no need to deny learning just because the system is unfamiliar. But acknowledging learning does not necessitate attributing experience. The next scientific step is to figure out the mechanism of history storage, the rules of transfer, and the limits of plasticity, and then compare them with the requirements of a particular theory of consciousness.

Another issue concerns the unit of analysis. In a distributed system, you can measure a separate area, an entire organism, or interaction with the environment it has changed. If traces of past activity are stored externally and influence future movement, part of the functional memory may be in interaction with the environment. Such a description requires an explicit indication of the boundary of the system, otherwise the external trace can easily be misrepresented as an internal memory.

Why the teaching debate requires replication

Markel in 2020 did not reproduce the previously stated effect of associative learning in peas. The work called into question the sustainability of a particular result. A negative attempt at reproduction does not prove the impossibility of all forms of learning in plants; a positive initial message does not resolve the question of qualia either. Evaluation requires comparability of conditions, control of alternatives and new independent verifications. [41]

In an association task, it is important to show that the result depends precisely on the connection between two events. If the group received more light, heat, or time to grow, the change may have another explanation. Control should, if possible, maintain the number of influences, changing their ratio. Otherwise, the researcher compares different physical conditions and attributes the entire result to communication learning.

Even a successful association check answers a limited question. In our own program, we can set a weight between two signals that increases when they appear together. The program will begin to predict the second according to the first. This is a clear functional result. To move from prediction to experienced expectation, additional justification will be required, applicable not only to the example you liked.

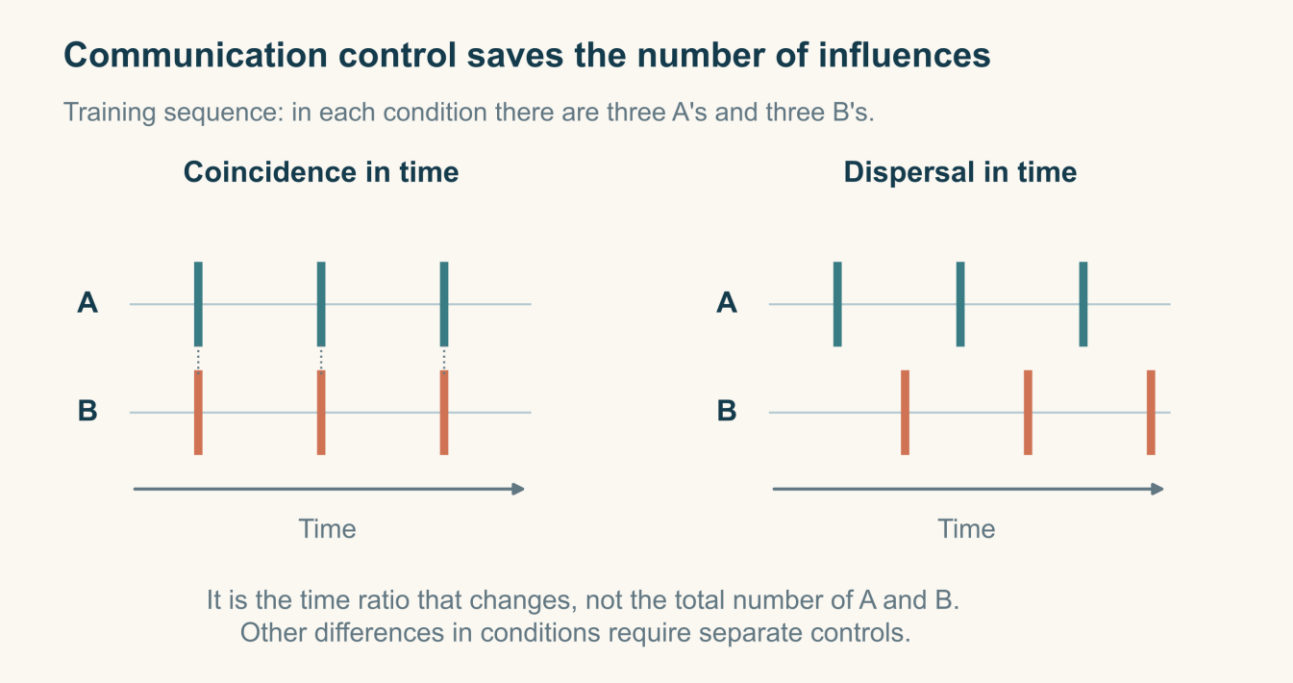

Figure 49. Division of the number of events and the relationship between them. The two fictional sequences have the same number of signals, but their temporal relationship is different. This control tests the role of coincidence while maintaining the number of influences; the shifted events can still form another predictable connection. To draw a conclusion about learning, it is necessary to test precisely the stated rule, and not to attribute the entire effect to differences in association conditions.

What should a hypothesis about experience add?

Suppose the researcher claims that any memory and any feedback is sufficient. It is necessary to apply such a rule to control devices, where both properties are implemented in a simple way. If the rule attributes experience to them as well, the author must acknowledge this consequence or clarify the original condition. You cannot save a broad formulation for an attractive biological case and subtly narrow it for an inconvenient engineering example.

Another hypothesis may require a special integration of cues and internal perspective. Then it should be shown how these properties are realized without neurons and how they differ from distributed regulation. The expression “an organism acts in its own interests” does not in itself define such a mechanism. Variables, dependencies, and conditions for breaking the function must be specified.

Finally, it is useful to distinguish between a positive hypothesis and the absence of a prohibition. Just because a theory does not exclude plant experience does not mean that it predicts it. For a positive conclusion, appropriate conditions must be established. This difference helps maintain research openness without unduly expanding the circle of confirmed bearers of experience.

Two models with the same beginning

In the habituation model, each signal has its own h, initially 0. The answer is 1 minus h, then h is increased by 0.2 the remaining distance to 1. In the depletion model there is a shared resource r, initially 1. The answer is r, then the resource is multiplied by 0.8. On the first five presentations of A, the responses of both models are equal to 1; 0.8; 0.64; 0.512; 0.4096. On the new B, the specific model corresponds to 1, and the total resource is already equal to 0.32768.

During each pause step, all h's are multiplied by 0.8 and r's are increased by 0.2 from missing to 1. First we write down the response, then we update the state. These are completely specified training rules, not measured parameters of the plant or slime mold.

What has now been established

A downward curve alone is not enough to select an explanation. A new signal tests specificity, a pause tests recovery. Both results must be calculated in the same predefined order.

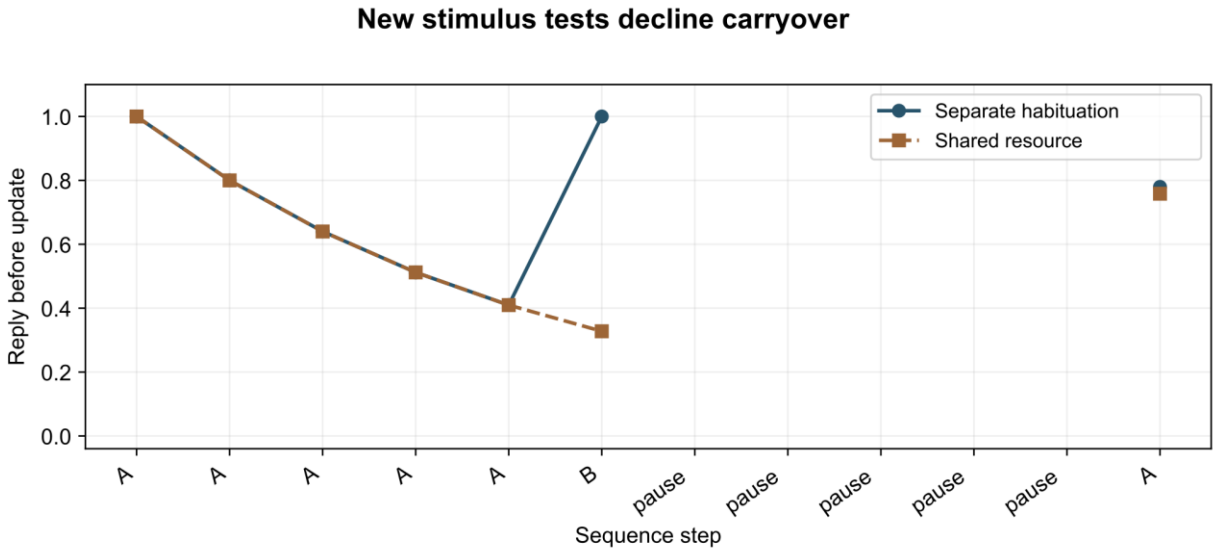

Figure 50. Habituation and exhaustion differ in transference. Calculation according to the explicit rules of the chapter: for the presented stimulus h ← h + 0.2(1 − h); shared resource after each event R ← 0.8R; during a pause h ← 0.8h and R ← R + 0.2(1 − R). The response is logged before the state changes. After five A, the first B gives 1 for a separate mechanism and 0.32768 for a resource mechanism. These are conditional mechanisms, not experience on the body.

Practice 22 Two models of one reaction reduction

Why do it: Separate specific habituation and a general resource using the same sequence of influences.

What you will need: State change rule from chapter 9 and control logic from chapter 7. Before practice, both algorithms are given in the same order below.

Time reference: 25–35 minutes. Can be done intermittently; This is not a speed standard.

1. Start with h for A and B equal to zero and a share of 1. Submit A five times, updating the state after each answer.
2. Predict the answer to B before calculating it. Then present B once.
3. Perform the five pause steps according to the rules above, then present A again. Compare the two models.
4. Check the calculations using python teaching_lab.py --chapter 22. Separately, write down what the result reveals about the mechanisms and what it does not establish about the experiment.

What to save: Step-by-step table of two models, new signal prediction, and specificity inference.

Hint: With new B, state h for A does not become state B. No cues are presented during the pause, so no answer is generated for these steps.

Research follow-up: Match the control to one primary work of the chapter. Indicate which observations the training model takes into account and which it does not represent at all.

Analysis and answers - Appendix F, chapter 22, p. 291

Self-test questions

1. Can functional memory exist without a nervous system?
2. Why doesn't a decrease in response prove learning?
3. Does the absence of neurons make the question of qualia meaningless?

Functional memory turned out to be wider than nervous tissue. The next chapter will consider artificial and biological implementations, for which it is important to establish the structure, boundaries and sources of self-description.

Chapter 23

Artificial and biological candidates

We looked at different organisms and saw how the material itself does not determine either the mechanism or the experience. Now let's move on to artificial and hybrid candidates. Before assessing their self-reports, it is necessary to describe the available inputs, the stored state, and the use of information. The basic practice in this chapter uses a ready-made minimal circuit. The full agent will be introduced later, in chapter 28; there is no need to contact it in advance.

Media question

Can experience occur in a program, robot, or grown neural system? The answer cannot be obtained from the title of the material. A program describes a way to organize calculations, a robot describes the connection of a calculation with a physical action, and a neural culture describes the biological composition of a certain system. None of these characteristics by itself defines the complete mechanism of consciousness.

Research requires distinguishing between medium, organization, interaction history, and available metrics. Two systems of identical cells can have different connections. One program can run on different hardware. The robot can be controlled by a local controller or receive commands from outside. If the boundary of the device in question is not specified, the properties of the entire installation can easily be mistakenly attributed to one part of it.

What can be deduced from the theory

Butlin and colleagues proposed evaluating artificial systems through features derived from scientific theories of consciousness, followed by testing their implementation. The article was published online in November 2025 and included in the journal issue 2026. The method is not a ready-made universal test: the conclusion depends on the validity of the selected theories, the correct translation of their requirements and the actual design of the system. [42]

The feature test must answer three questions. Is it necessary, sufficient, or merely incidental according to this version of the theory? Which part of the device implements it? What intervention can be used to show that it actually fulfills its intended role? If a sign exists only in the developer's description, this is a reason to study the architecture, and not count the condition as fulfilled.

In chapters 13–15 we looked at different areas of requirements. The accessibility of selected content to multiple processes is important to the workspace. For the recurrent approach, a certain organization of reverse interactions is essential. For higher-order theories, a representation of one's own state. For approaches sensitive to intrinsic physical causation, functional similarity may not be sufficient. These differences should be preserved when assessing a candidate, and not combined into an indifferent sum of “signs of reasonableness.” [16] [17] [19] [20]

The program and its self-description

Let’s imagine a program that, after any failure, displays the phrase “I’m confused.” Another version of the same program displays "error code 7". If replacing the text does not change the calculation, the message itself does not prove the presence of experienced confusion. It is part of observable behavior, the origin of which must be investigated. Conversely, the absence of an emotional vocabulary does not establish the absence of all potentially relevant mechanisms.

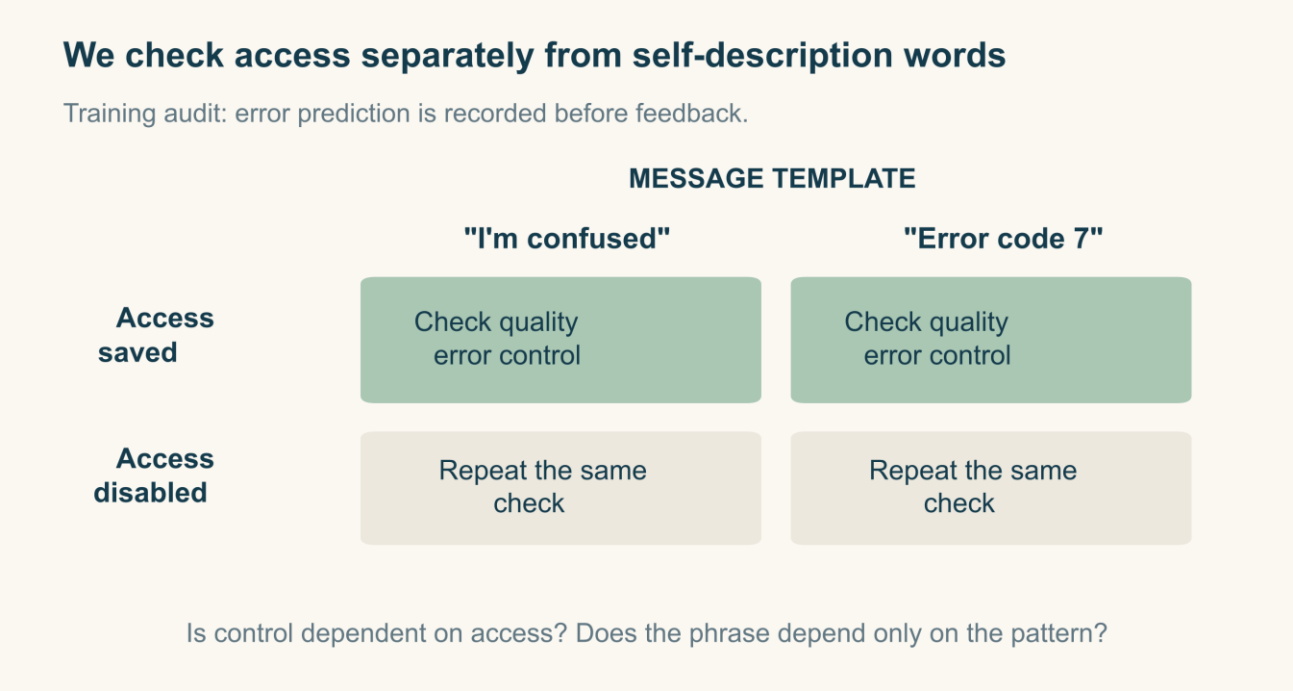

Figure 51. Separate audit of information and self-description. The design independently changes the program's access to internal information and the wording of the message. If control depends on the first factor and words on the second, their causal roles differ; This is a preliminary conditional prediction, and not the result of an experiment already performed. It is necessary to exclude access to the hidden correct answer from the environment.

A more insightful candidate is able to predict its own error before obtaining a result, consider memory limits, and change strategy when a systematic failure is detected. This can already be checked functionally. However, the story about the reasons for the decision can be formed separately from the decision itself. Then a

plausible explanation does not guarantee access to the reasons that were actually at work.

In your own experiment, it is useful to separately change access to internal information and the response pattern. If the quality of control depends on the internal state, and textual explanations depend only on the template, the functions differ. If the program uses a hidden correct answer from the environment, successful assurance may generally be a data leak. Therefore, auditing inputs is as important as analyzing received phrases.



Figure 52. Artificial system in a physical environment. Subject illustration of a mobile device with sensors and computing equipment. When evaluating such a candidate, it is necessary to establish the boundaries of the system: where the command is formed, what information is available to the controller, and what consequences of the action it receives.

What does the body add?

The robot allows the interaction to be closed: an action changes the world, a new observation depends on the action, and the model can clarify the relationship between command and effect. Such a device is convenient for studying self-reference and adaptation. It does not make corporeality a sufficient condition for consciousness. It is necessary to establish exactly which functions of the physical body cannot be replaced without loss by another organization of interaction.

Consider an educational robot with a camera and a moving platform. Rotation changes the image. If the device predicts a change, it can distinguish between the

movement of an object and the effect of its own rotation. If there is a subtle change in control, you will need to update the action model. In this example, the body specifies specific causal dependencies, rather than a symbolic right to use the word “I.”

If a theory asserts the necessity of a particular bodily regulation, testing it requires specifying which variable is being regulated and why it is significant. Battery reserve, represented by a number, does not automatically become an experienced need. Its integration with choice, learning, memory, and the proposed principle of connection with experience needs to be explored. Without this, we will only rename the usual control signal.

Neuronal cultures, organoids and hybrids

Neuronal culture represents biological nerve cells in artificially organized conditions. An organoid is a grown three-dimensional system that models certain aspects of tissue development and organization; this name does not mean a ready-made brain. The hybrid connects a biological component with computational control, sensors or actuators. Detailed experimental examples and limitations will be discussed in chapter 31.

The same principle of analysis is important for our purpose: learning, electrical activity, and successful control should be measured separately from perceived experience. Biological origin makes some mechanistic hypotheses more meaningful, but does not automatically inform their implementation. It is necessary to clarify the structure of connections, feedback boundaries, the role of the external algorithm and the temporal stability of interactions.

Particularly indicative is a hybrid in which an external controller selects rewards and converts tissue responses into commands. Changing the result of the entire installation may occur due to the controller settings. To attribute learning to biology requires appropriate control comparisons. To attribute experience, a separate theoretical argument will also be required. These are two sequential questions that should not be combined with one project name.

System class	What should you describe first?	Common error in drawing a conclusion
Program	Real states and dependencies of calculations	Consider self-description as a direct measurement of experience
Robot	The connection between the perception of action and the state of the body	Taking the presence of a hull as a sufficient condition

System class	What should you describe first?	Common error in drawing a conclusion
Neural culture	Organization of the cellular network and external control	Attribute the properties of the entire installation to the network
Organoid	Degree of organization and available interactions	Identify the tissue model with the whole brain
Hybrid	Boundaries of parts and distribution of causal functions	Do not differentiate between learning the controller and the biological component

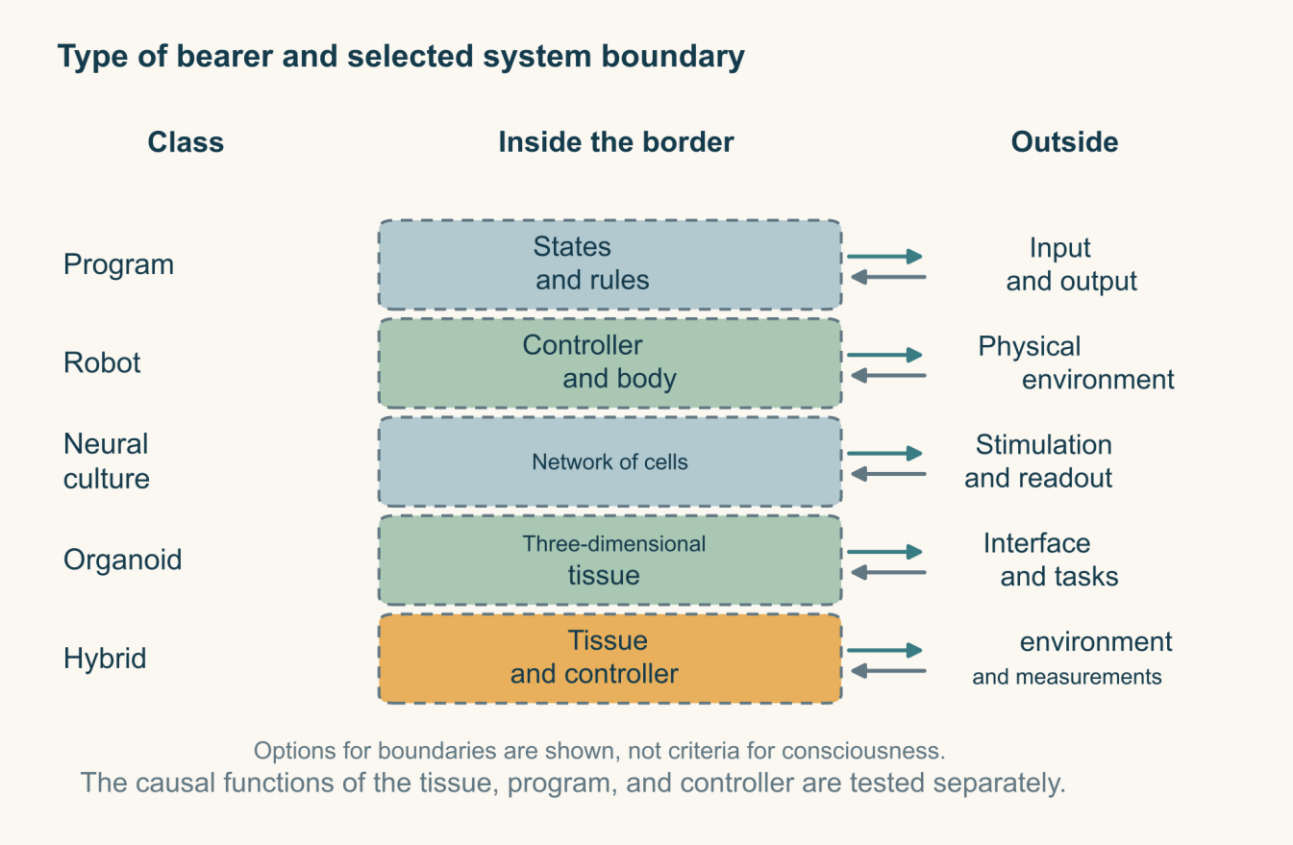

Figure 53. The medium does not define the boundary of the study. Five examples of system boundary selection. The dotted line separates the part in question from the external interaction. For a program, robot, culture, organoid, and hybrid, it is necessary to establish where learning, state storage, and control are performed. These are options for describing the installation, and not a mandatory structure of all systems of the corresponding class and not a rating of the probability of consciousness.

How not to turn signs into ratings

Let's imagine two candidates. One has well-tested error control, but the organization of content distribution is unknown. The second has a common exchange channel, but the assessment of its own reliability is formed according to

a ready-made rule. Simply adding up the ticks will hide the difference. For one theory, the absence of a certain feature is critical; for another, the same feature is not necessary.

Therefore, it is better to present the audit result as a profile. For each requirement, the theory, interpretation, measurement, outcome, and remaining alternative are specified. Unknown properties are noted separately. If hardware is not available, a description of the program does not replace knowledge of physical causation when it is essential to the approach being tested.

Also check whether the intended feature is actually being used. The module can be connected to shared memory and never read the content of interest. Another module is capable of storing a confidence score that does not influence any choice. The presence of fields in the code then exaggerates the functional organization. Selective deletion, distortion, and restoration of the condition help establish a causal role.

However, the absence of change after removal does not always give a final negative conclusion. Perhaps the task did not require this function or it was performed by a duplicate mechanism. It is necessary to provide in advance the condition where the function should be required, and check this explanation of compensation. The architectural audit then becomes an examination of the working device.

A Minimal Design before the Full Practical Course

Let the device store the right target probability p and the direct control probability q. The selector module reads both numbers and selects a command. The monitor reads the same numbers and calculates the expected success. A text report can either be generated from the monitor or inserted according to a fixed template. Changing the template while the states are saved should not change the command selection. A distortion of q, on the contrary, can change both the choice and the forecast. This is a causal testing scheme; Full update rules will be discussed in chapter 28. It is already possible to distinguish an architectural requirement from the impression created by the text.

What has now been established

We can create a profile even with incomplete knowledge of the implementation, if the unknown properties are designated as such. The presence of memory, body frame or neural cells is not a substitute for testing specific connections.

Practice 23 Candidate's specification

Why do it: Separately evaluate access to states and verbal self-description on a finished diagram.

What you will need: World model and action model from chapter 15; internal access and external journal from chapter 14. The name of a component is not a test of its implementation.

Time reference: 25–35 minutes. Can be done intermittently; This is not a speed standard.

1. Draw four components: estimates p and q, choice, monitor, report. Make only the connections indicated in the analysis.
2. For each connection, indicate the sender, the recipient, and the consequences of disconnecting it. Do not add a hidden true target to the monitor.
3. Compare two interventions: changing the phrase pattern and distorting q with the same external input. Record expected changes to selection, forecast, and text separately.
4. Compile a profile on two theories: what the circuit models, what has not been tested, and what basis for drawing conclusions about the experiment is still needed.

What to keep: Mind map and profile of checked, unchecked and missing properties.

Hint: The figure does not imply that all components are actually implemented. When you go to the code, the status of each connection will need to be confirmed.

Research continuation: After chapter 28, repeat the profile on the active agent and his log. Do not replace the implementation check with a block name match.

Analysis and answers - Appendix F, chapter 23, p. 291

Self-test questions

1. Why is successful conversation about experiences not sufficient to draw conclusions about them?
2. Does a biological bearer guarantee consciousness?
3. Why is a candidate’s profile better than the number of characteristics completed?

The assessment of other candidates prepared the language of the requirements. Now let's choose a limited explained difference and build our own model, which can be compared with meaningful competitors.

Part 7

Developing your own explanations

Chapter 24

Requirements for the future model

So far we have looked at the available explanations and implementations. Now let's start our own project. Consider the specific question of the influence of source history on affiliation assessments. It is narrow enough to construct a mechanism and at the same time connected with the broader theme of the subject. First, we determine that the model must explain what facts have already been published and what results are still only our expectation.

From an interesting question to an obligation to explain

Imagine that you are controlling the image of a hand on the screen. The movement usually follows your command, but is sometimes delayed. After several minutes of stable communication, a short delay seems to be a random interruption. If the connection was constantly changing, the same delay may be perceived differently. A substantive question arises: does experienced belonging depend only on the current match of cues or also on the history of the putative source? This is a mental teaching example, not a report of an experiment. Its purpose is to transform the general idea of self-continuity into specific requirements for the model.

The model begins with a commitment to explain difference. The phrase “the brain builds itself” allows for almost any outcome. The statement “given the same current alignment, the history of a stable source changes the subsequent dynamics of membership” already limits expectations. Now we can ask how long the influence of a story lasts, whether it transfers between sources, and whether it persists when the mode of response changes. The more precisely these questions are posed, the clearer what result will be the success of the explanation.

Identify the object with the required accuracy

The initial subject of the project is the experience of bodily belonging. In the experiment by Kalckert and Ehrsson, the ownership of the artificial hand and the sense of authorship of the movement could diverge. Therefore, belonging cannot be determined solely by successful arbitrary control. The model must account for this difference by maintaining separate indicators of ownership, control, and evaluation of the correctness of the action. [4]

The subject can be written in one sentence: “We explain the change in experienced belonging when the history of consistent bodily signals changes.” The next line specifies the scope of application: adult participants, short episodes, a specific method of presentation and reporting. This record retains the possibility of expansion. It does not require now to explain color, dreams, biographical self-awareness and the initial emergence of consciousness in one variable.

When extended, the explanatory obligation itself will change. First, the model predicts body-ownership ratings. Then you will need to justify their connection with the experiencing organization. Finally, a stronger hypothesis should say why the proposed organization constitutes or supports the presence of experience. These transitions will become separate tasks, rather than a subtle increase in the volume of a single output.

Distinguish between data and desired results

It is useful to have two lists. The first contains published results that the model must take into account. The second is the yet untested expectations of our project. If you mix them up, your own assumption begins to look like an established fact. For example, a statement about the transfer of history between two similar images still belongs to the second list, if we have not found such a comparison in the literature.

The work of Samad and colleagues shows how belonging can be examined through a common cause inference model of visual, proprioceptive, and tactile signals. This is the closest published guideline, not our new idea. One’s own contribution will have to be sought in certain dynamics, transfer conditions or the discrimination of alternative mechanisms. [15]

Each entry in the first list must include an objective, a metric, and an limits of inference. "Signals combine" is too broad. “Under such conditions, the assessment of belonging changed, and at the same time the position of the hand was measured” already allows us to find out what kind of observation is being reproduced. Also note whether the results of several indicators are consistent. Their divergence may be the main requirement for a future model.

Four requirements for our project

The first requirement concerns the present. The model must respond to the current compatibility of the signals: their coordination cannot be completely insignificant. In this case, it is necessary to determine which compatibility features are available to the system. Geometric coincidence, temporal proximity and dependence on action are different information; they cannot be replaced by the undeciphered “consistency” variable.

The second requirement concerns the past. The influence of history must have computable dynamics. It is not enough to add memory and then declare any delay to be its confirmation. It is necessary to indicate in advance what is preserved: individual events, an assessment of the stability of the environment, the expectation of a common cause, or the history of a specific source. Each solution offers its own mechanism of forgetting and transference.

The third requirement concerns identity. If a story belongs to a specific source, the model must explain why it is returned when it appears again. At the same time, a competitor is needed that tracks external objects in the same general way. Otherwise, the advantage of source memory will be mistakenly called evidence of a special mechanism of self-awareness.

The fourth requirement concerns measurement. The same internal state can give different answers when the wording of the question or the criterion of agreement changes. Therefore, the assessment generation model must be separated from the source model. This separation will allow us to check exactly where the effect arose: in the processing of signals, in their attribution, or in the way the result is reported.

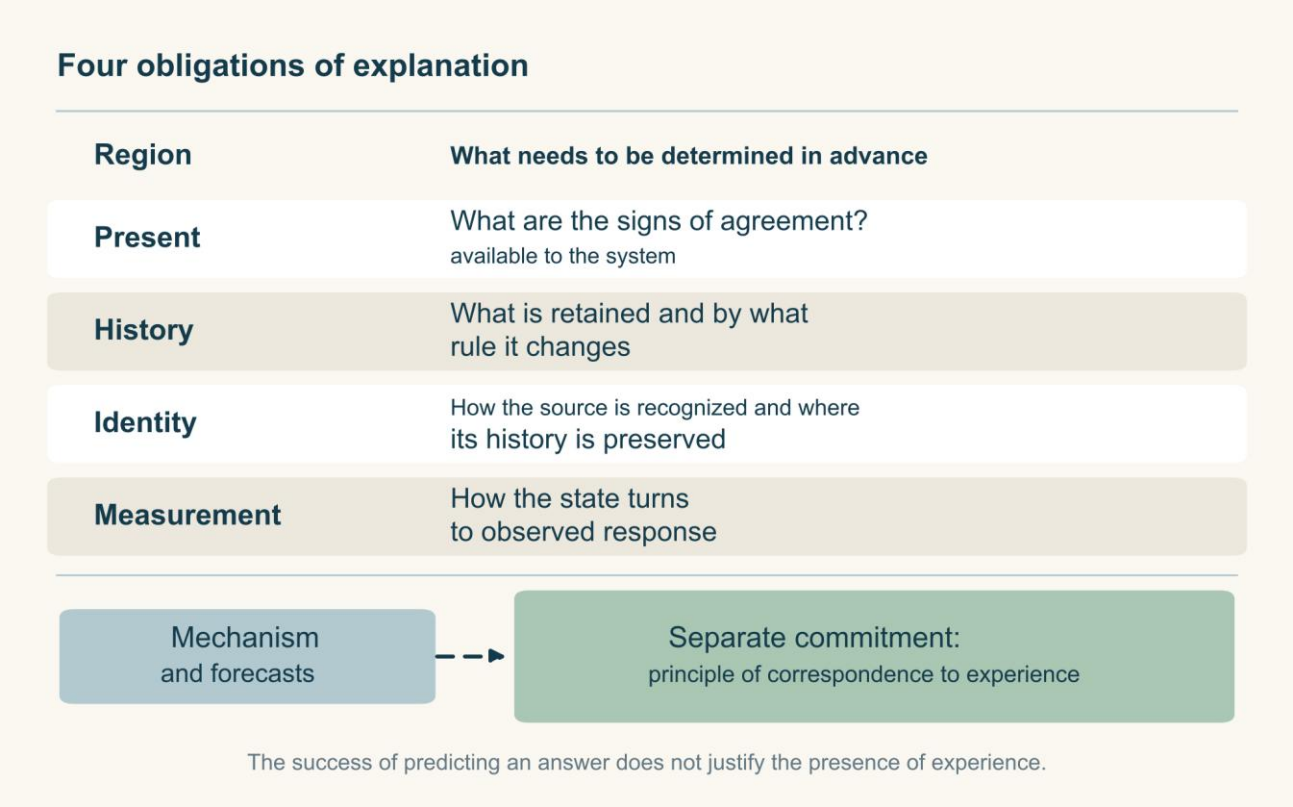

Figure 54. Model commitment map. Methodological scheme. Four requirements relate to design and model verification. The dotted line denotes a separate argumentative task: to explain the connection between the mechanism and experience; it does not mark a fixed transition from accurate forecast to the presence of experience.

Formulate several competitors

Let's call M0 the current coordination model. It uses only the selected current window. M1 maintains an overall history of confirmations and contradictions without separate records of sources. M2 maintains distinguishable histories of suspected sources. These names represent families of future computing; For now, specific rules need to be chosen for each family.

Let's add M1I - a general model for tracking the identity of objects. It is needed because M2 can benefit from conventional thread splitting. A comparison of M2 with M1I should show whether additional bodily organization is required or whether a mechanism applicable to any stable object is sufficient. A possible answer of “a fairly general mechanism” would be the result of an investigation into the nature of self-reference.

Don't intentionally make your competitor weak. If M0 receives less current information, and M2 additionally sees hidden labels, their comparison checks the advantage of the entry. If M2 has significantly more configurable parameters, it needs to be checked on new data. The requirement for fair comparison is included in the model specification from the very beginning.

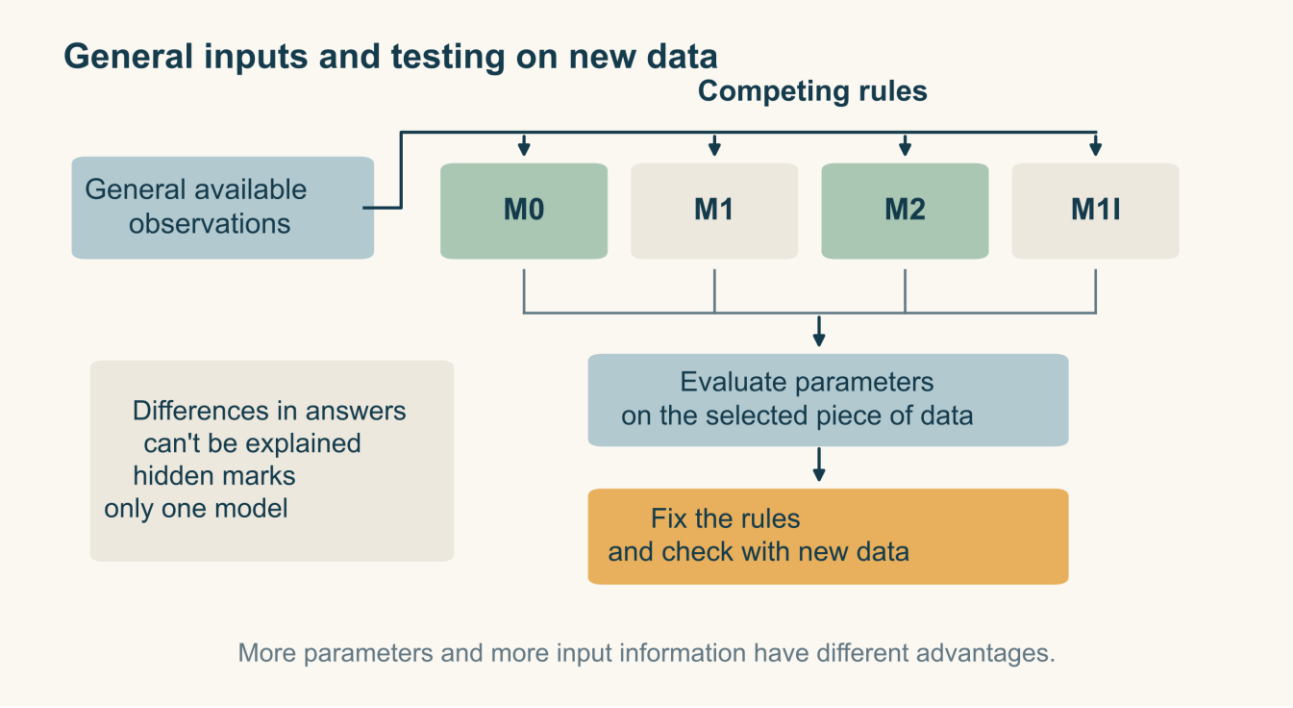

Figure 55. A fair comparison of computing competitors. M0, M1, M2 and M1I are compared with a common set of available observations, preserving the differences in their memory rules. Options are evaluated separately from testing on new data, and the hidden source label cannot be provided only to the preferred option. The advantage in such a comparison relates to the given task and does not yet confirm the special nature of self-reference.

Set a positive criterion for success

Suppose the model were to predict the restoration of affiliation upon the return of a familiar source. Success will not be simply matching the average score. We need a predicted direction, rate of change, and conditions under which the effect disappears. It is especially valuable if the parameters estimated in one problem help predict another without new arbitrary settings.

For a training project, you can set a requirement: the model must distinguish between two sequences with the same number of consistent events, but different organization of sources. This is not yet an experimental pattern. It is the chosen test that challenges the theory to explain the role of order. In the next chapter we'll build a simple sequence where the computational options actually diverge.

It is useful to supplement the success criterion with a stopping condition. If the available data does not distinguish between M1 and M2, the project does not yet answer the question about the history of the source. Then you should change the task, rather than increase confidence in your favorite model. A good specification helps you see in time what research needs to be done next.

How a shared vision turns into a commitment

The idea of “history helps you feel yourself” allows for too many results. A clarification relates the initial difference to the procedure: given the same test sequence, the membership assessment should be restored differently after returning a familiar or new source. Next, you need to indicate what familiarity means, what sign the source recognizes, and how the dynamics of the response are measured. A competitor is recorded separately, which stores the history of any external objects. If he gives the same prediction, the special bodily function has not yet been highlighted. Such a limitation helps to choose the next experiment, rather than devaluing the project.

What has now been established

A good requirement limits both the expected success and the conditions for revision. It is now possible to create a specification in which past, present, identity, and measurement are given separate roles.

Practice 24 Make requirements for one model

Why do it: Prepare requirements for one model before choosing convenient calculation rules.

What you will need: Distinguishing between ownership and authorship; common cause model; fair comparison of competitors. The success of predicting the answer and the presence of experience require different reasons.

Time reference: 35–45 minutes. Can be done intermittently; This is not a speed standard.

1. In one sentence, define the difference being explained and the scope of the conditions. For basic work, use the return of a source after another source has intervened.
2. Separate published limitations from your own expectations. For a published result, indicate exactly what was measured.
3. Write down four requirements: current signs, preservation of history, recognition of identity, formation of a response. Name M0, M1, M2 and the general tracker as future competitors.
4. Determine the result that would support the chosen difference and the result that would make the special version redundant. Don't assign arbitrary human averages just yet.

What to keep: One specification with the subject, requirements, competitors and criteria for the next decision.

Tip: Increasing the number of parameters and adding available information have different benefits. Both need to be checked.

Exploratory continuation: Propose a second problem to which the model should transfer parameters without free tuning. Explain why she tests the same relationship.

Analysis and answers - Appendix F, chapter 24, p. 292

Self-test

1. Why does not the retention of information about the past in itself confirm a special mechanism of self-reference?
2. What does the advantage of M2 show over M0 if the first one receives a hidden source label, and the second one does not?
3. Why separate the established limit and your own expectation?

The requirements do not yet determine the next step of the model. In chapter 25 we will set the input, state and update and follow one calculation to the end.

Chapter 25

From idea to mechanism

In chapter 24 we required an explanation of the influence of the history of a certain source. Now you need to define a procedure that, based on the old state and the new observation, unambiguously calculates the next state. This will be the first executable part of your own project. We'll look at updating one probability in detail, and then change only the memory organization: no carryover, general record, and separate source records.

The mechanism must allow the next step

In the previous chapter, we required that the model explain the influence of the history of a particular source. Now we need to make the idea feasible. If you tell the model the current state and a new observation, it should determine the next state. The phrase “the system integrates important information” does not provide this: it is not known what is considered important, how the integration occurs, and what is stored between events.

Let's start with a small design. It does not depict the structure of the brain in detail. Its task is to show how different rules for handling history give different predictions. All numbers and sequences in this chapter are educational. They are chosen for transparent calculation and do not describe measured human experiences.

Separate input, state and output

An event of two types is received as input: one that supports the general cause of the signals or one that contradicts it. In a real project, such information would have to be obtained from time, position and connection with the action. Here we will temporarily replace the multidimensional event with a binary feature. This is a clear simplification, the scope of which will need to be checked.

The latent state of the model is the probability that the current signals are related to a common cause. Let's call it a belief. This is a technical word: the model stores a number and updates it according to a specified rule. It does not add an internal observer who understands the meaning of the number. The output for now represents the renewed belief itself. The connection with the human assessment of belonging will be a separate transformation.

Finally, distinguish between environmental truth and model estimation. The experimenter can know which events are actually related. The model receives only

imperfect features. If you give it a true causal label, the inference task disappears. Therefore, each input should be checked by asking: in what observable way could the system obtain this information?

Build one update rule

Let a supporting event occur with probability 0.75 for a common cause and with probability 0.25 for different causes. This is our chosen feature reliability model. With the initial belief 0.5, one supporting event gives rise to a new belief 0.75. A contradictory event with the same initial state gives 0.25.

Between events we will weaken the old belief in the direction of uncertainty. The training version retains eight tenths of the deviation from 0.5. Therefore, the state 0.75 before the next observation turns into 0.70. Relaxation can be interpreted as allowing the underlying situation to change. This is the chosen dynamics of the model, not an established law of memory.

```
def update(previous, supports_common_cause, retention=0.8):
    prior = 0.5 + retention * (previous - 0.5)
    likelihood_common = 0.75 if supports_common_cause else 0.25
    likelihood_separate = 1.0 - likelihood_common
    common_weight = prior * likelihood_common
    separate_weight = (1.0 - prior) * likelihood_separate
    return common_weight / (common_weight + separate_weight)
```

This fragment defines the transition between states. It does not compute qualia. It can be done manually, on a calculator, or in Python. Each parameter has a separate role: the reliability of a feature determines the strength of new evidence, preservation determines the influence of the past. If the result changes, you can now determine which rule caused the change.

One transition in general form

Let us denote the retained probability of the common cause by the letter p, the retention coefficient by r, and the reliability of the positive attribute by a. For the basic example, r = 0.8 and a = 0.75. Before observation, we obtain p_do = 0.5 + r × (p − 0.5). The value r = 0 completely removes the previous deviation, r = 1 retains it.

For a positive attribute, the weight of the common cause is equal to p_do × a, the weight of different causes is equal to (1 − p_do) × (1 − a). The new p is obtained by dividing the first weight by their sum. For a negative sign of likelihood, the positions are swapped. This choice assumes symmetrical reliability of the two outcomes and is necessarily included in the description of the model.

There are three different numbers in the calculation. The stored p refers to the previous processing; p_do takes into account a possible change in the cause; the new p takes into account the already obtained feature. To predict the sign itself, we use one more value: P(+) = a × p_to + (1 − a) × (1 − p_to). It is this that is recorded before observation, when the quality of the forecast is later assessed.

The numbers of the future sequence should not appear in the rule. Even if the environment is known to the developer, the model receives only the allowed features and the label of the observed source. When we later make identity recognition more difficult, the unmistakable label will be replaced by a separate score rather than added as a hidden clue.

For M0, M1 and M2, only the entry selection for this update changes. Therefore, the differences can be traced back to a specific memory rule. This is much stronger than the general claim that history is important: we now know the cases of agreement, divergence, and the conditions under which the resulting effect should disappear.

Three ways to use one update

M0 before each event starts with 0.5. Thus, the supporting event always results in 0.75. M1 maintains one common belief for the entire sequence. M2 stores a separate belief for each distinguished source. When a source appears, the model updates the record associated with it. In this simplified version, relaxation occurs on the next record access rather than at each calendar step.

Consider three events: source A supports the common cause; source B contradicts it; source A supports her again. For M1 the states will successively be 0.75, then 0.4375 and approximately 0.7105. For M2, we get 0.75 in record A, 0.25 in record B, and 0.875 after returning A. M0 will produce 0.75, 0.25, and 0.75, respectively.

The difference arose from the treatment of history. When A returns, model M2 uses its past confirmation; M1 also carries a recent contradiction from B. We have received a specific prediction that can be verified by calculation. But for now this is a prediction of the internal dynamics of our design. Linking it to human data will require a measurement model and empirical validation.

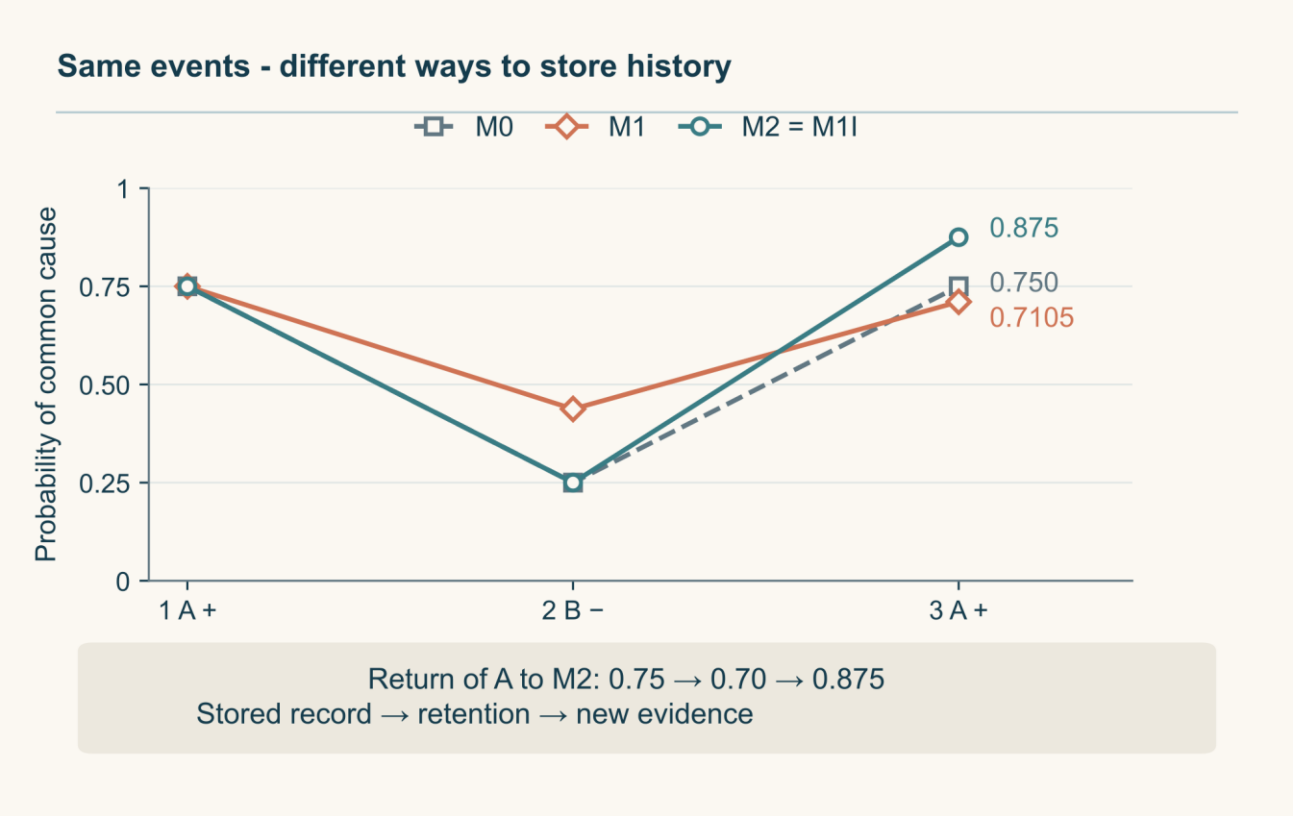

Figure 57. Return to source history. Accurate training calculation for the sequence A+, B−, A+. Reliability of the trait is 0.75; a hold of 0.8 is applied before the corresponding record is updated. M0 starts with 0.5, M1 carries over the general history, M2 and M1I keep the same separate records. These are states of computational models, not measured experiences or evaluations of participants.

Where does the source identity come from?

The most dangerous workaround is to pass M2 an error-free source name and then assign it the ability to discover identity on its own. In educational calculations, the letters A and B stand for idealized discrimination. In a complete design, it is necessary to explain its origin: for example, the source is recognized by a stable combination of position, appearance and expected consequences of the action.

Such recognition may also be uncertain. If an event with probability 0.7 belongs to A and with probability 0.3 to B, you can distribute the update between records or store several possible histories. These solutions are not equivalent. The first smoothes out differences, the second preserves competing versions. The choice should be based on predictions and computational limitations rather than on name convenience.

Then you will need a common competitor M1I from the previous chapter. It uses the same identity features for external objects. If M1I explains the data just as well, an additional special mechanism of bodily self-reference is not yet required.

The mechanism can be general in structure and at the same time participate in the content of self-awareness.

Add a Measurement without Hiding the Explanation in It

Suppose a participant rates the sense of ownership on a scale from zero to ten. The simplest teaching relationship multiplies the model's belief by ten. The final event in the sequence would then give M0 a value of 7.5, M1 about 7.1, and M2 a value of 8.75. These are conditional predictions of a response, rather than experimental results or a quantitative measure of experience itself.

A more realistic relationship might include individual response criterion, scale nonlinearity, and random variability. But each added element increases the flexibility of the explanation. If any internal state is allowed to be transformed into any response, it becomes impossible to compare mechanisms. Therefore, the method of generating the response must be determined in advance and checked whether it transfers between conditions.

It is useful to design an additional task that helps assess the reliability of sensory discrimination regardless of ownership. Then the model will not simultaneously guess the quality of the input, the strength of the story, and the response criterion using one number. Multiple data sources may constrain different parameters, although definitive identifiability requires separate verification.

Check edge cases

With save set to zero, each update starts at 0.5. The differences between the use of past beliefs disappear. With a persistence of one, the old state is carried over without weakening. If events cease to distinguish between common and distinct causes, a new observation should not change the belief. These cases follow from the chosen rules and serve as implementation checks.

Then change the order of events while maintaining their number. Rearrange the names A and B everywhere at the same time: the result should change only in the names of the records. If renaming itself changes the membership prediction, the program probably has an unnecessary rule. Such a test tests notation invariance rather than the existence of subjective perspective.

Don't limit yourself to a pretty sequence. Find conditions where the models coincide. If the source is always the same, M1 and the simplified M2 behave the same. Therefore, an experiment with only one source will not be able to test the value of individual stories, no matter how many observations it contains.

Why the Same Evidence Can Yield Different Estimates

Before A reappears, its stored probability is 0.75. Holding 0.8 retains eight-tenths of the deviation from 0.5: the prior probability becomes 0.70. A positive sign has a likelihood of 0.75 for a common cause and 0.25 for different ones. The resulting weights are 0.525 and 0.075. Their sum 0.600 is needed for normalization, so the new probability of the common cause is equal to 0.875.

This probability is not equal to the prediction of the next positive sign. To make a forecast, you also need to take into account the unreliability of the observation: 0.875 multiplied by 0.75 plus 0.125 multiplied by 0.25, that is, 0.6875, if the hidden cause did not change until the next observation. If a cause change is allowed, the corresponding transition is applied first. These two quantities cannot be discreetly assessed using the same outcome.

What has now been established

New numbers emerged from explicitly chosen rules. Now you can check which saved record is involved in the update and what will happen when it is replaced.

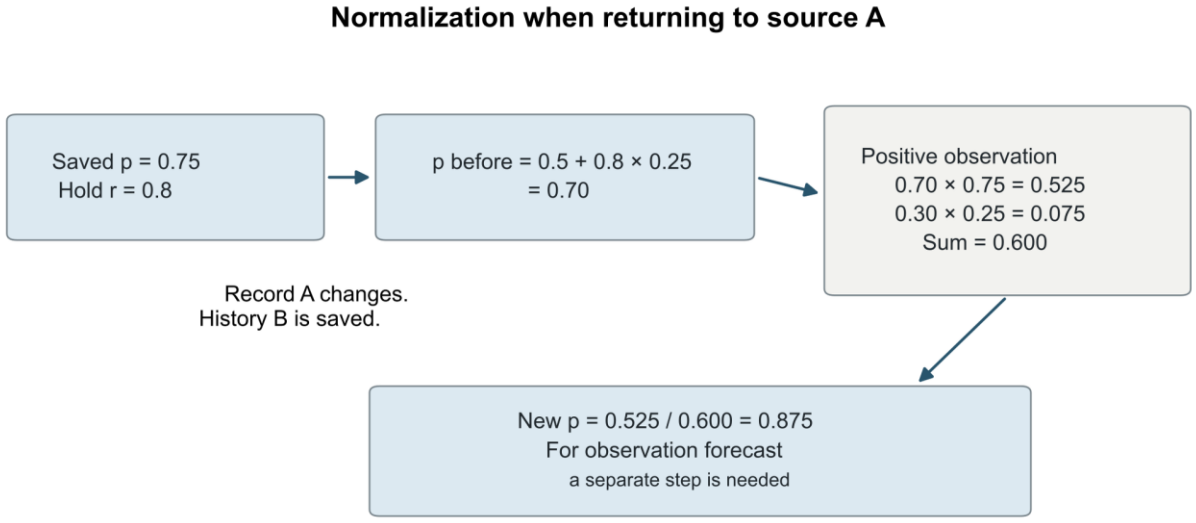

Figure 56. Normalization of one update. Full head step 25 with reliability 0.75. The preliminary state after retention, unnormalized weights and the final probability are labeled separately. The probability of a hidden cause does not replace the prediction of an observed symptom.

Practice 25 Trace the causal chain

Why do it: Recover four events for three memory options and a common tracker.

What you will need: Normalization of probabilities and the difference between observation and hidden cause from chapter 11. History hold here is applied when accessing a record; calendar time is discussed separately in chapter 33.

Time reference: 35–45 minutes. Can be done intermittently; This is not a speed standard.

1. Use the sequence A positive, B negative, A positive, B positive. Everywhere the initial belief is 0.5; reliability 0.75; hold 0.8.
2. For M0, start with 0.5 each time. For M1, transfer one record. For M2, update only the recognized source entry. The shared tracker uses the same individual entries.
3. At each step, write down the selected record, the state after holding, the two weights and the normalization result. First, calculate it yourself, then open the analysis.
4. Repeat holding zero and check the renaming of A and B throughout the sequence.

What to save: Four step-by-step calculations and two controls with a previously known result.

Hint: M2 returns negative history B at the fourth step. Separate memory is not required to provide the greatest confidence.

Exploratory follow-up: Replace error-free source labels with observable attributes. Describe how recognition uncertainty will change record selection, and compare a strong general tracker on the same inputs.

Analysis and answers - Appendix F, chapter 25, p. 292

Self-test

1. Why does M2 support common cause more strongly at the third event of the learning sequence?
2. What would be lost by using only one source for the entire study?
3. Why is the transformation of belief into affiliation assessment a separate part of the model?

The memory mechanism has become executable. Now it is necessary to explicitly formulate a stronger statement: which organization, according to the author's hypothesis, could be associated with quality and experience.

Chapter 26

From mechanism to experience

We can calculate the history of causes and predict certain responses of the model. The central hypothesis of this chapter adds an independent step: it involves the connection of some physically realized mutual organization with a minimum of experience. Let's look at this step one by one. First we define the mechanism under study without phenomenal words, then a strong statement of sufficiency, then structural consequences and methods of objection. This order preserves the research's courage and makes its commitments negotiable.

What question has not yet been answered?

The model from the previous chapter can store history, distinguish between sources, and change the probability of a common cause. Let's say it predicts human affiliation ratings well. We will learn something significant about the mechanism of these estimates. However, the question remains: why are the corresponding processes accompanied by experience at all? In Chalmers's philosophical setting, an explanation of the ability to discriminate and communicate is not automatically an explanation of the presence of subjective experience. Here we use this distinction as a requirement of the argument rather than as a pre-established prohibition on explanation. [2]

To move forward, we need a correspondence principle: an explicit statement of how the system's organization relates to the quality and presence of experience. It may turn out to be wrong. But if there is no such statement at all, the technical project will never go beyond what is directly programmed into it. Let's formulate a strong working hypothesis and find out what obligations it creates.

Define organization without the word experience

We propose to study processes in which several distinctions mutually determine each other. One component distinguishes the properties of the observed object. The second evaluates the source of the change. The third specifies what additional observations or internal transformations are now possible. Their states are updated together: the inferred source changes the attribute's interpretation, the attribute changes the source's expectation, and the selected test rebuilds both representations.

Reciprocity here has a causal meaning. For each hypothesized relationship, targeted interventions and expected changes in both directions are specified in advance: for example, how a change in the source's evaluation will change the

interpretation of the attribute and how a change in the presented attribute will change the source's evaluation. General activity or correlation is not enough. In addition, the system must distinguish between alternatives that are not currently occurring: what would change with another observation, another command, or abandonment of the previous interpretation. These alternatives should influence processing and not exist only in the developer's explanation.

Finally, individual distinctions form a general dynamic. It is not possible to replace each component with an independent reproduction of its normal outputs and store responses to all predefined interventions. This check distinguishes mutual determination from the coincidence of multiple recorded streams. It does not define consciousness by name; it specifies the system device under study.

A Working Correspondence Principle

Let us formulate a hypothesis: a certain class of physically implemented processes of mutual determination of differences is sufficient for minimal subjective experience; the qualitative organization of this experience corresponds to stable relations between the distinctions and alternatives available to the process. This is the author's research hypothesis for this book. Its independence relative to existing approaches and its truth have yet to be established.

In the working version, the class is limited by three requirements: several separately variable features or degrees of freedom of content are presented; their changes are causally connected in closed dynamics; the relationships between them are used to select and review alternatives. Separate variability does not mean statistical independence: it is the interaction of features that is the subject of verification. A counter, history record, or self-name is not enough to meet these requirements. However, the presence of all three is also not yet an established sign of experience: it is the sufficiency of the organization that constitutes the strong part of the hypothesis.

Sufficiency does not mean necessity: the principle does not say that without this organization any experience is impossible. Predictions of qualitative changes should follow from its structural part and the conditions of a specific intervention.

In principle, no verbal "I" or constant external movement is required. The choice may concern internal refinement of the representation. Self-ownership appears as additional content when the system associates changes with a sustainable source of its own impact. Therefore M2 can become part of the proposed mechanism of self-awareness, but the table of individual stories itself does not yet implement the entire proposed organization.

Why is this hypothesis worth testing?

The first reason is the unity of explanation. Mutual constraints can simultaneously specify which contents are compatible, how they are preserved, and why changing one changes the others. Instead of separate, unexplained labels for quality, unity, and time, a single, explorable process emerges. This is the merit of the design, not proof of its truth.

The second basis has to do with perspective. The state matters for subsequent changes to the system itself: it determines what differences it can check and what errors it can correct. An external observer can describe this process, but its causal organization is not created by description. The hypothesis proposes to consider such internal certainty as a possible basis for subjective perspective.

The third basis is the ability to derive consequences beyond the original problem. If the proposed organization is related to experience, breaking certain reciprocal connections should change its organization in a predictable way, even when the simple distinction remains. Reconciling such predictions with independent human data would provide grounds for considering the implementation as a candidate. The transition to an artificial system will still require an argument about the transferability of the principle.

Quality through relationships and invariants

Let the learning system differentiate between hue, brightness, and source of visual change. The value of one coordinate in its vector is not a ready-made “redness”. The relationships need to be explored: which states it mixes, which it contrasts, how changes in luminance affect hue discrimination, and what interventions maintain the difference. The resulting structure is much more meaningful than a separate numeric code.

If we only renamed states or changed coordinates, retaining all significant transitions, the working hypothesis predicts the preservation of high-quality organization. If the salience relationship and their mutual influence have changed, it predicts a change in organization. This is how an invariant arises: the content depends on certain relationships, and not on the notations chosen by the developer.

This does not yet establish which human color experience corresponds to the machine state. For such a comparison, additional grounds are needed. Even complete coincidence of the structure of similarities does not automatically resolve the issue of absolute qualitative correspondence. At the same time, the explanation of the relationships between qualities represents an independent meaningful result and limits the possible future answer.

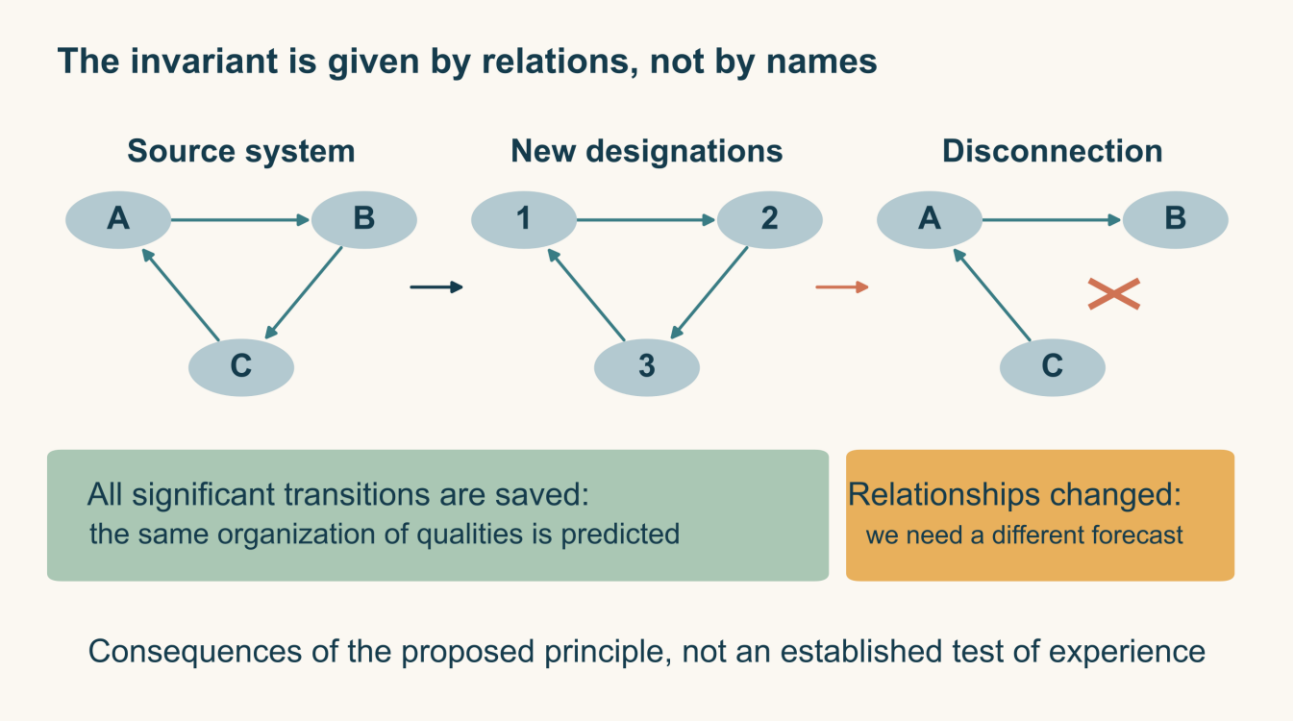

Figure 58. Renaming and changing the causal structure. A schematic example of an invariant of the working hypothesis of chapter 26: renaming preserves significant relationships, and removing a relationship changes the subject of the forecast. The principle offers sufficient organization for the experience; the necessary disappearance of experience does not follow from the disruption of a separate connection.

Explaining the Existence of Experience

The strong part of the principle asserts the sufficiency of the organization, and not just the consistency of its structure with self-reports. We take this as a testable identification hypothesis in a broad research sense: some process of mutual determination and minimal experience represent one specific process described from different perspectives. Such an identification cannot be achieved by simply defining the word “experience.”

Its justification must consist of independent connections: explanations of already available experiences, predictions of their changes, elimination of competing mechanisms and consistent application to new realizations. If the theory only explains the texts of the machine, the foundation remains weak. If one mechanism links a variety of phenomenal differences to specific causal changes, the argument becomes stronger, although the logical possibility of alternatives remains.

The critic has the right to ask why mutual determination should be felt at all. Our principle does not derive the answer from neutral arithmetic: it proposes a concrete identity and takes on its consequences. Future work should show that this identity provides better explanation and prediction than its competitors. In this

way we really pose the question of the existence of experience, without replacing it with the success of one function.

Clarify implementation and boundaries

The word “physically” does not automatically resolve the issue: any running code has a medium. It is required to establish what causal organization should exist in the execution itself. If only a certain computational structure is essential, various suitable devices can implement it. If additional physical interactions are significant, program matching is not sufficient. These versions of the principle must be separated before the results can be compared.

IIT 4.0 provides an important example of a theory that explicitly relates proposed properties of experience to physical postulates. Our workpiece is not an IIT calculation and does not derive its basis from the similarity of the words “reason” or “unity”. The comparison must address specific requirements and divergent implications. [19]

The example of a simple regulator is especially dangerous. It can have feedback and several states. Therefore, it is necessary to check whether our three requirements really exclude overly simple designs, and on what basis. If an exception has to be introduced just because we don't like the result, the principle is underdefined. This counterexample helps improve the definition of the class under study.

Three stages of testing a strong hypothesis

The first stage establishes the organization. We need several distinguishable states, certain rules for their mutual change, and the participation of alternatives in further processing. For each suspected connection, an intervention is specified. Having a correlation, a common source, or a nice loop in the diagram is not enough. The result of this stage can be positive regardless of the acceptance of the hypothesis about the experience.

The second stage connects the organization with quality relationships. Let the theory assume that the content belongs jointly to the source. It should then predict how interference with source evaluation will change the discrimination of ambiguous contents and which relationships will remain the same. We need to compare this with an alternative that changes only the response criterion while maintaining the experience. The usual correctness of classification does not solve this issue.

The third stage concerns the presence of minimal experience and transfer between implementations. Here the strong principle proposes the identity or sufficiency of a

certain class of processes. The argument is supported by a system of independent correspondences and predictions, an analysis of alternatives, and a consistent application of the boundary. An organizational test from the first stage does not become an independent proof of sufficiency just because it is repeated many times.

This arrangement allows progress without the promise of a final solution with one program. One can rigorously demonstrate the desired causal reciprocity, discover an incorrect structural consequence, or establish that a proposed concept is too broad. Each result changes the research program in a particular way.

Computable example of two directions

To test the logic, we will build another proprietary training circuit. Let p and q be between zero and one. They denote the evaluation of an ambiguous attribute and the evaluation of a source. Let us introduce the function s(z) = 1 / (1 + exp(−z)), which transfers any real number to this interval. Both new estimates are calculated from the previous ones simultaneously: p_new = s(x + b × (2q − 1)), q_new = s(h + d × (2p − 1)). Here x and h are the available external evidence, b and d are forces in two directions.

With x = h = 0, b = d = 1 and the original p = q = 0.5, the estimates remain 0.5. Now, with the input unchanged, we will force q = 0.8 before updating. The next p will be approximately 0.645656. A separate intervention p = 0.8 changes the next q by the same amount. If b = 0, the first direction disappears; if d = 0, the second one disappears. Thus, the reciprocity condition receives a specific verification matrix.

The check selection option can be set separately: the query is enabled when 1 − |2p − 1| exceeds the selected price. This is a conditional policy used for the sake of transparency. It shows how changing the view can change the next step. The verbal name of the request does not add experience to it.

The command python teaching_lab.py --chapter 26 reproduces the baseline condition and targeted interventions. The example tests the causal organization of certain numbers. It does not yet realize the strong demand for quality structure and minimum experience. To do this, it is necessary to determine the content of the distinctions, derive the corresponding phenomenal consequences and justify the physical implementation.

A strong alternative controller should receive the same observations and comparable resources. If it saves the entire answer matrix, the selected test does not yet highlight our organization. This is a meaningful possibility of a result. The conclusion about uniqueness cannot be obtained by defeating only a pre-recorded fixed sequence.

Testing Reciprocity without Defining Experience by the Test Result

Let p evaluate an ambiguous attribute, q its source. In the training loop, the source estimate affects the next p, and the feature estimate affects the next q. If the external input remains unchanged, let us forcibly change q. If p changes according to a given rule, one direction is checked. Then we independently change p and check q. One of these experiments is not enough to make a statement about two directions.

Even the success of both interventions so far establishes a causal organization. The correspondence principle additionally presupposes its relation to experience. For a qualitative conclusion, it is necessary to state in advance which relationships between contents will change and which will remain. To transfer to another medium, you need to clarify which device should actually exist in physical form. Therefore, the next practice begins with an audit of a ready-made principle, and an independent proposal becomes the second level of work.

What has now been established

We separated three results: the presence of a reciprocal relationship, the prediction of difference structure, and the strong statement of the presence of experience. They can be part of the same project, but require different reasons.

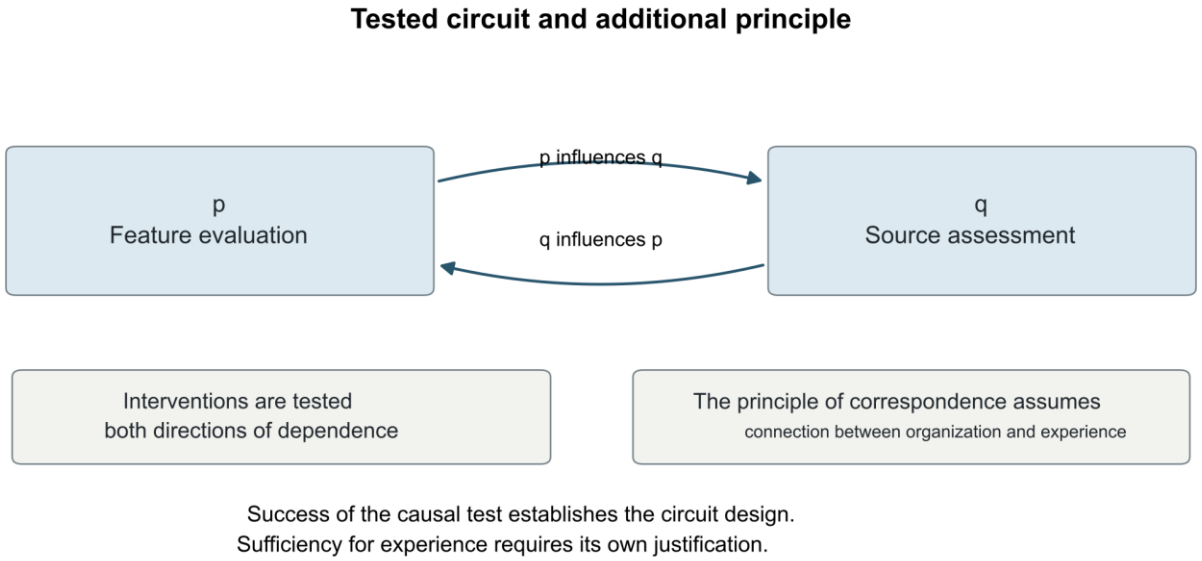

Figure 59. From causal organization to the hypothesis of experience. The teaching loop p and q is implemented in teaching_lab.py --chapter 26 and tested by two targeted interventions with one-way controls. The connection of such an organization with experience is a separate author’s hypothesis. Violation of the proposed sufficient condition does not prove the absence of any experience.

Practice 26 Protect one principle

Why do it: Test the obligations of the correspondence principle and distinguish sufficiency from necessity.

What will be needed: Causal intervention, sufficiency and necessity; separate story from chapter 25. First, read the full analysis of the circuit and principle, then move on to your own version.

Time reference: 40–60 minutes. Can be done intermittently; This is not a speed standard.

1. Rewrite the principle of the chapter in three sentences: organization, intended connection to experience, structural consequence for qualities.
2. For each requirement, provide an observation or intervention to verify the fulfillment of the organizational condition without the words “consciousness” and “experience.”
3. Consider three controls: independent output entries, a one-way controller, and a simple regulator. Indicate which conditions of the principle they satisfy and which are unknown.
4. For one intervention, separately write down the functional prognosis and the phenomenal consequence that still requires justification. Do not infer the absence of any experience from the violation of a sufficient condition.

What to keep: The principle card, three counterexamples, and two separate levels of predictions.

Hint: The phrase “this process is sufficient” allows for the possibility of other sufficient processes. The absence of a chosen process is therefore not a logical negation of all experience.

Research continuation: Offer your own updated version. Derive the invariant, the change in the essential relation, and the revision condition. Compare it with the closest scientific approaches before declaring the independence of the idea.

Analysis and answers - Appendix F, chapter 26, p. 293

Self-test

1. How does a strong principle in this chapter differ from good report prediction?
2. Why shouldn't renaming coordinates itself change the predicted quality?
3. Can the proposed hypothesis be considered proven after assembling a system with the three specified properties?

The principle received separate obligations. The next chapter will turn the differences between engines into a design that can support one version and limit another.

Chapter 27

Competitors and risky predictions

A working hypothesis is useful when a possible outcome can change its content. Now let's prepare a distinguishing project. The first stage concerns the history of sources and projections of estimates; following mutual causal relationships and qualitative organization. Let us not demand that one comparison solve all the issues. For each stage, we will determine our own indicator, a strong opponent and a review condition.

Difference must be dangerous to the idea

Prediction is useful when a possible outcome can change our attitude towards the model. The statement “memory sometimes influences decisions” is easy to reconcile with almost any observation. The statement “when source A returns, a certain history will change the trajectory of the estimate more than the history of another source, given the same current signal” creates a risk. One might obtain a different trajectory, no transfer, or an effect entirely explained by general learning.

In this chapter, we will prepare an experimental design. It has not yet been carried out. His numerical examples refer to the training model, and future human responses remain unknown. This formulation allows you to make a forecast before the temptation arises to adjust the explanation to the result.

Start with matching models

M2 from chapter 25 stores separate histories of A and B. The general identification tracker M1 is capable of doing the same. If it receives the same features and uses the same rules, the two models are computationally the same. No sample size will differentiate between identical predictions. One must either accept that a general mechanism has been proposed, or define an additional substantive difference.

This is not a failure of the project. A general cause-tracking mechanism may be involved in bodily self-reference. It would be a mistake to declare it specifically corporeal just because the problem depicts a hand. The first stage of the study is entitled to answer a narrower question: is a separate source history required to explain the measured estimates?

For a strong hypothesis from chapter 26, the following step will be needed. What matters here is the participation of history in the general mutual dynamics of perception and choice of verification. Simply storing sources and including them in such a circuit are different organizations. They must be compared on the basis of

their previously inferred consequences, maintaining the distinction between functional outcome and phenomenal inference.

Design four conditions

The proposed human experiment uses common visual and motor tasks with a virtual pointer or hand image. The study will require a professionally designed protocol and voluntary participation; This is where its logic is developed. The participant first interacts with source A when the signal connection is stable or variable. Then the same test sequence from the previous or new source B is presented.

Four combinations are obtained: stable history and former source; sustainable history and new source; changing history and former source; changing history and a new source. The total duration, number of events, and characteristics of the test sequence are compared. The source differs on a preselected observable attribute rather than on the participant's affiliation response.

It is important not to define sustainability solely by the number of matches. Two sequences can have the same number of confirmations and contradictions, but a different order. The model must receive the entire real sequences and produce a time forecast. If the original regimes differ in several characteristics, each of them is included in the list of possible explanations.

Measure several different results

The main result of the first stage will be the pre-selected dynamics of membership estimates. The manner in which scores are obtained should differ minimally between conditions. Additionally, current delay discrimination, source recognition, action attribution, and confidence can be measured. These indicators will help to understand what process has changed, but should not become unchecked into new key outcomes.

There is no external right number for a given person's experienced belonging. Therefore, the quality of sensory discrimination cannot be used as an automatic check of the truth of the report. It answers another question: how accessible is the physical difference that the model considers an input to the participant? Source recognition also helps verify that participants received the information required by the M2 mechanism.

Preserve the wording of the questions and the order of measurements. If control is assessed before ownership, the first rating may change how the second is made. This possibility within the design should be investigated in a pilot, rather than

assumed to be absent. The pilot helps refine the procedure; the confirmatory test then follows a separately fixed plan.

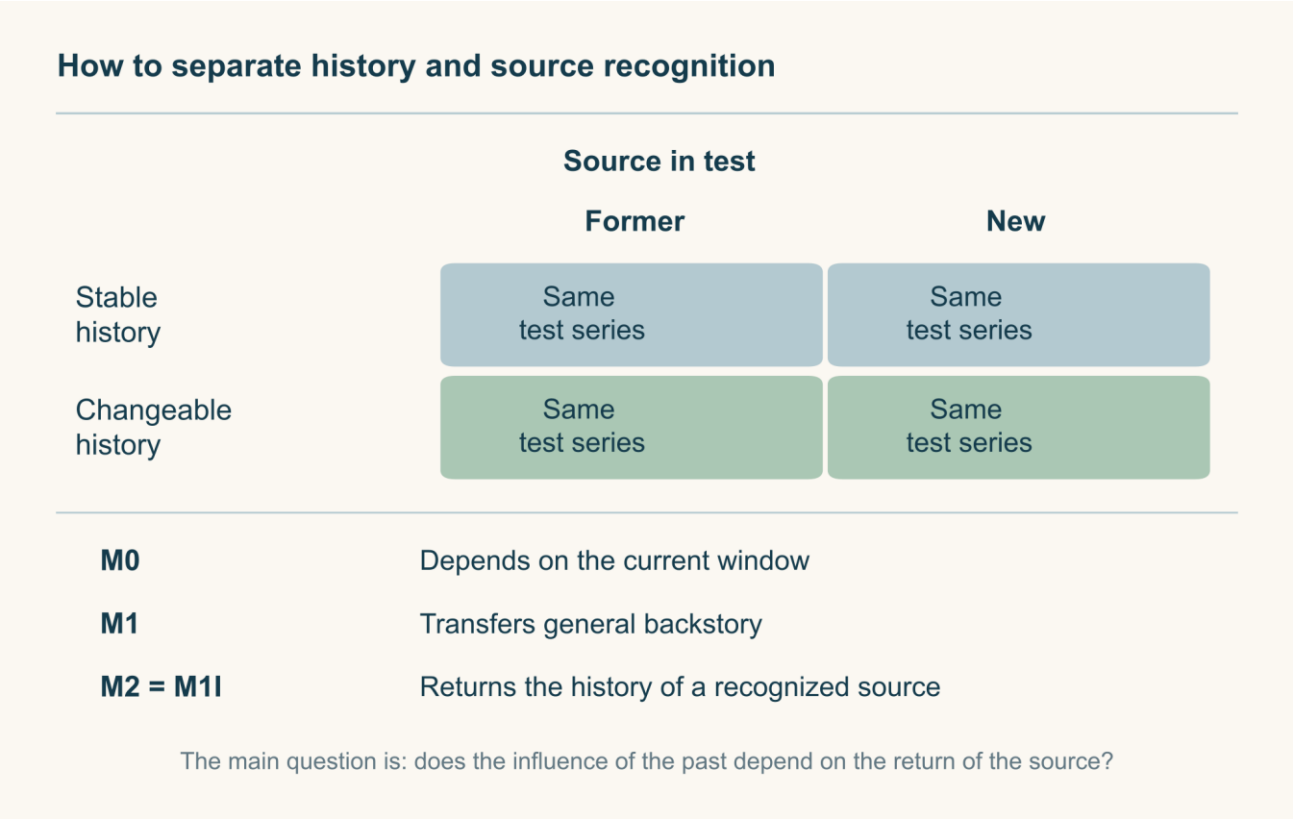

Figure 60. Four conditions of a discriminating design. The project has not yet been completed. Two factors - the nature of the background and the return of the previous or the emergence of a new source - are combined in four conditions with a comparable test series. Below are the qualitative differences of the rules, not the expected human averages. In the selected implementation, M2 and the general tracker M1 are the same; this design alone does not differentiate them.

Get specific forecasts

M0 predicts the same internal state given the same current window. M1 allows for the influence of backstory, but carries it over without separating the sources. M2 predicts the return of the corresponding individual story. Therefore, the main differentiating effect is the dependence of the influence of the past on whether the previous source has returned. Direction and magnitude must be calculated from fully defined rules.

The training sequence A is positive, B is negative, A positive gives M0 the last state 0.75, M1 approximately 0.7105, M2 - 0.875. These values are not expected human averages without additional assumptions. They show that differences between mechanisms can be obtained before the results are observed. The response pattern and individual parameters will require separate assessment.

The common tracker M1I in this sequence coincides with M2. Therefore, the gain over M0 and M1 will support the benefit of taking into account the history of the source in the explanation under consideration, but not yet the special physical nature of the mechanism. This conclusion requires transfer between the bodily task and the comparable task of tracking an external object.

Check mutual definition

For the next stage, we will create two software implementations. The first uses source estimation to select additional observations; the test result changes perception, history, and next choice. The second reproduces the usual outputs of the components, but breaks one of these feedback connections. In normal sequence, their behavior may appear similar.

The intervention must make the difference noticeable. For example, let’s change the reliability of one feature after generating the history of the source. The reciprocal system must rearrange its test choice according to its changed uncertainty. Playing back a recorded stream will not provide this. A comparison with a full-fledged alternative controller is also necessary so that defeating the record does not become the only achievement.

Such an experiment tests the presence of a certain causal organization. The structural part of the hypothesis of chapter 26 additionally requires predicting which relationships between experienced contents will change with the chosen intervention. Testing will require independent experiential data and a transference argument. From the mere sufficiency of the organization it does not follow that its rupture excludes any experience. Identical human texts at the output of the program do not provide a separate phenomenal basis.

Compare using new data

If the parameters are selected over the entire sequence, its successful description is not yet a successful prediction. Divide the data: estimate the parameters on one part, compare the predictions on the other. When there are multiple participants or biological samples, separation must take into account their structure. Random mixing of closely related events can make checking too easy.

Computational research also requires testing whether the chosen procedure can distinguish between models when the model that generated the data is known. Wilson and Collins view parameter and model recovery as important tests of computational analysis. In our project, such checks will be carried out on our own artificial sequences. [43]

It is advisable to look for a strong difference where rivals disagree, and not where everyone has an easy time. However, selecting conditions after reviewing the underlying human results should not be presented as a pre-determined success. This choice can become a useful research discovery and the basis for the next independent experiment.

Write down the terms of the revision

If M1 predicts new estimates just as well, a separate source history is not yet justified. If M2 wins, but M1I fully explains the advantage, the special body version is redundant. If the advantage occurs only after free reconfiguration of the response transformation in each condition, it is unclear whether the mechanism or flexibility of the measurement part explains it.

For a strong hypothesis, the absence of the phenomenal changes predicted by its structural component for a specific intervention would pose a distinct problem. We must determine whether the intervention succeeded and whether an alternative pathway was operating. Yet we cannot indefinitely dismiss every inconvenient result as an insufficiently genuine disruption. The criterion for a successful intervention must be fixed before testing.

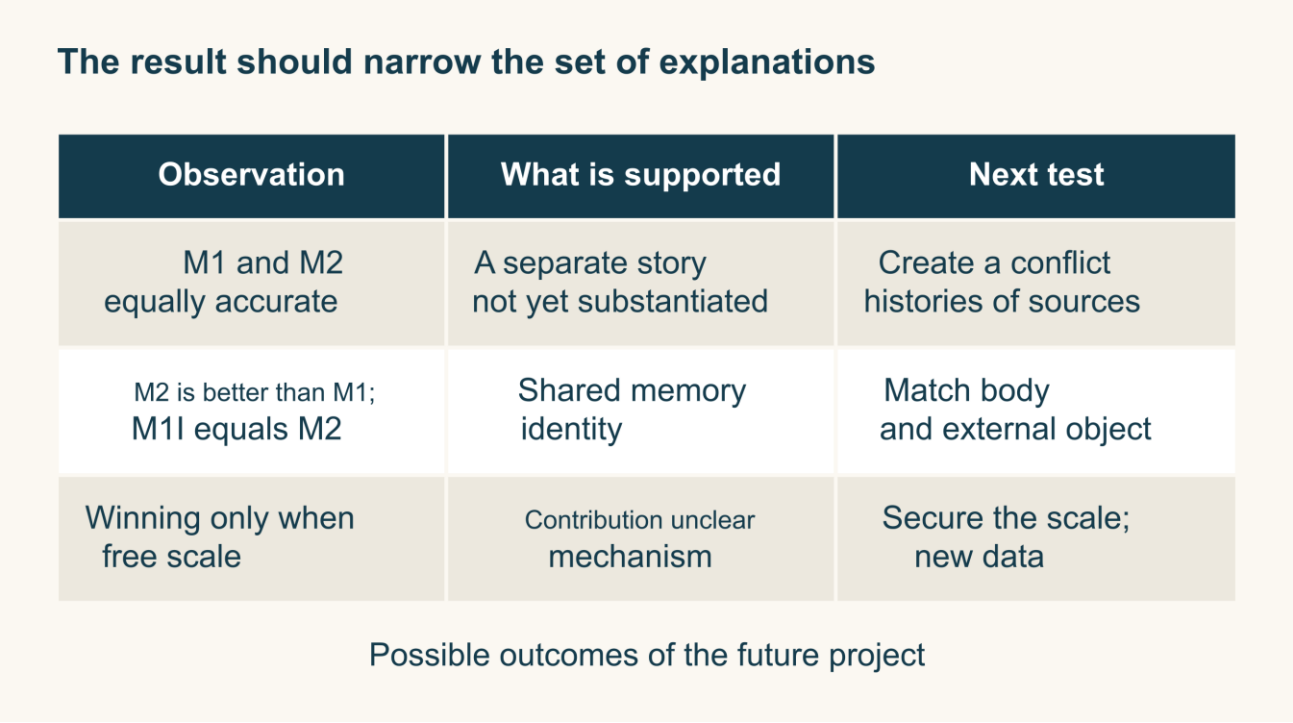

Figure 61. From possible outcome to discriminative test. The table specifies three possible outcomes of a future comparison and different decisions after them. The coincidence of M2 with the common tracker M1 leaves the question of special self-reference open, even if separate memory is useful.

Why won't more data distinguish the same rules?

M2 and the general tracker store separate histories and, in a minimal implementation, update them in the same way. If you give them the same observations and initial states, they will calculate the same predictions for any length of such experiment. Increasing the sample will reduce random bias in data estimates but will not introduce missing differences between models. Therefore, the first design might test the utility of a single story relative to general memory, but a specific bodily explanation would require an additional mechanism or other comparison.

What has now been established

The discriminative result is determined by the predictions, not by the name of the model. Now let’s write out the conditions and conclusions before the main data appears.

Practice 27 Make a differentiating project

Why do it: Prepare the four conditions of the first stage and separately identify the next causal test.

What you will need: M0, M1, M2 and a general tracker from chapters 24–25; the correspondence principle from Chapter 26. The human experiment here is designed, not considered already completed.

Time reference: 35–50 minutes. Can be done intermittently; This is not a speed standard.

1. Make four combinations: stable or changeable history and the same or new source in the test. Specify the available features and the same test sequence.
2. For each condition, write down which story M0, M1 and M2 are used. Keep the general tracker under your own control, even if it coincides with M2.
3. Name the main result and additional indicators: do not combine affiliation, source recognition, authorship and sensory discrimination with one number.
4. Set one targeted intervention for the next step and control for its side effects. Save the result and revise your preferred version.

What to keep: A design with four conditions, distinguishable predictions, and an explicit boundary to the first conclusion.

Hint: The numbers 0.75, 0.710526 and 0.875 from the calculation do not become human averages without a model for measuring and estimating the parameters.

Exploratory follow-up: Prepare two implementations with the same normal outputs but different directional influences. Compare also a full-fledged alternative controller and check the transfer to new scenarios.

Self-test

1. Is it possible to increase the sample to distinguish between the computationally identical M2 and M1I?
2. Why is beating playback of recorded outputs a limited outcome?
3. What to do with an interesting effect noticed after analyzing the basic data?

The project specifies what needs to be checked. Now let's take an active agent, whose inputs, memory, prediction and action are visible, and reproduce functional checks in code.

Part 8

From words to practice

Chapter 28

An Agent in an Environment with a Verifiable Task

We are moving from design to an operational program. It will evaluate the target, learn its own control, and select additional surveillance based on its cost. All these functions have already been prepared in the previous chapters. The first pass can be done without changing the code: understand the problem, write down the prediction and reproduce the finished run. Own new conditions appear after checking the original result.

The first agent on a regular computer

Open the workshop directory and find the agent_lab.py file. We have before us a small agent who chooses an action under conditions of incomplete information. All his work is available for viewing: what information was received, what was retained in memory, how the assumption changed and what forecast was made before the action. This is a convenient starting point for designing more complex systems. We can change one mechanism and observe the consequences without relying on the verisimilitude of the program's verbal response.

You don't need to understand every line of Python to get started. First, analyze the problem and predict the result. Then run the finished program, compare the expectation with the numbers, and only then proceed to implementation. The code here serves as an accurate record of the model. It can be read gradually: inputs, status, update, action, feedback. These five questions will help you understand even an unfamiliar program.

The installation app provides commands for Windows, macOS, and Linux. Requires Python version 3.10 or later. The program uses the standard library, runs locally and does not require an API. In the directory with the file, run the command below. On Windows, replace python3 with py -3 if your installation uses this launcher.

```
python3 agent_lab.py --steps 1000 --runs 30 --seed 1729 --json my_results.json
```

The steps parameter sets the number of steps of one script, runs — the number of repetitions, seed — the initial value of the randomness generator. The last

parameter specifies the result file. With the same parameters, the program repeats the same computational experiment. This helps test code changes, but other scenarios are needed to evaluate robustness.

How the environment works

At each step there is a target to the left or to the right. The agent does not see it directly. Between adjacent steps, the side is preserved with probability 0.95. Therefore, recent information is usually useful, although sometimes it is outdated. The usual observation correctly indicates the side in 75 percent of cases and is missing in a quarter of the steps. For a small price you can get additional observation with an accuracy of 90 percent.

The agent sends a left or right command. In the first half of the scenario it usually causes an effect on the corresponding side, and in the second half the connection is inverted. A single motor error with probability 0.05 also changes the effect. This creates two different sources of uncertainty: where the target is and what will happen after the command. Reliable goal perception will not help if the control model is outdated.

The rule changes from step 501 at a thousand steps. The experimenter knows this. The agent receives neither a switch flag, nor a schedule, nor a true target, nor a seed. After the command, a visible effect is available to him. Success information is calculated by the environment and returned to the researcher after the script completes.

Why not transfer success to the agent immediately? In a given binary problem, the combination of a known effect and success would reveal the true goal. Then memory and noisy observation would be tested in a different information situation. The prohibition relates to a specific research model. In another problem, correctness feedback may be a legitimate part of the interface, as in the strategy review work in chapter 30.

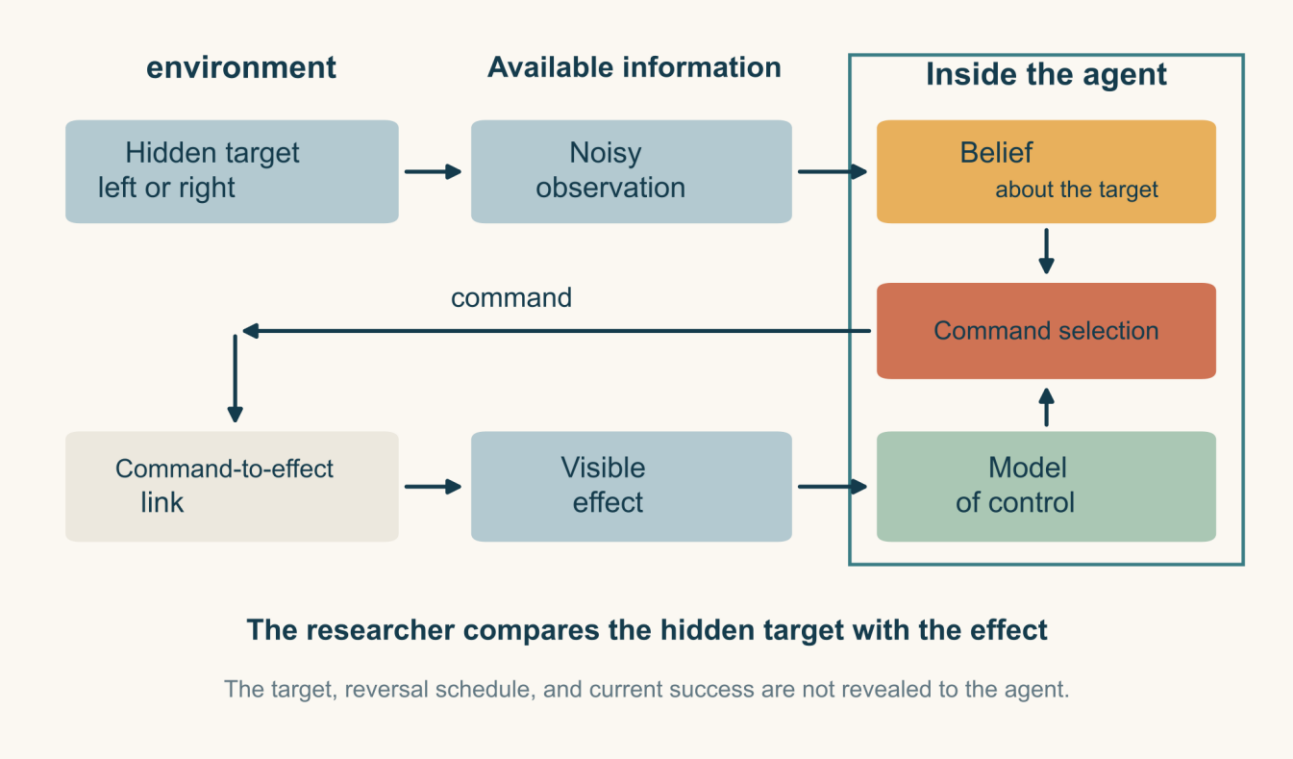

Figure 62. Two circuits of one agent. Training diagram of the interface of the first work. Observation updates the belief about the goal, and the visible effect of the command is the control model. The true goal, the moment of the rule change, and the current success are not included in the information available to the agent. An additional paid request provides another observation of the target and is not shown separately for clarity.

Memory in the form of probabilistic belief

The target_right_probability state stores the probability of the right target according to the current model. The number 0.8 means that, given the given premises and observations, the right side is considered more likely. This is a model statement that may be wrong. The true position is not included in the calculation.

Before making a new observation, the agent takes into account the possibility of the target moving. If the probability of the right side was 0.8, and the target persists with probability 0.95, the following prediction is obtained from two possibilities: the target was on the right and stayed there, or was on the left and moved. Their probabilities add up. In this example, the new forecast is 0.77. Observation then refines it using Bayes' rule.

```
old_belief = 0.8
stay = 0.95
prior = stay * old_belief + (1 - stay) * (1 - old_belief)
print(prior)  # 0.77
```

The memory here does not represent a detailed autobiography. The past is compressed into a single number, the sufficiency of which is determined by the

chosen probabilistic model. This storage method allows the causal contribution of history to be clearly shown. If an observation is missed, the agent still has a prediction; if history is disabled, the initial belief about the goal reverts to 0.5.

The second variable, direct_mapping_probability, estimates the likelihood of a command being directly associated with an effect. After the action, the agent compares the command with the visible result. The discrepancy may be due to a motor error or a change in control. Bayesian updating changes their relative plausibility. The model allows for a small probability of a link change at each step and is therefore able to revise a previously reliable assumption.

Why measure success before taking action?

From the two beliefs, the probability of success for each command can be calculated. The agent combines options for the target position and possible effect, selects the more promising command, and writes down a prediction before acting. If you record confidence after receiving the correct answer, the indicator no longer reflects previous uncertainty.

An additional observation is purchased when the expected improvement of the current solution exceeds the price. The program considers both possible sensor responses, recalculates the belief for each, and averages the best expected result. This is a calculation of immediate benefit. The future value of information in the next steps is not separately planned, so we do not call such a policy optimal for all possible horizons.

Cost is needed to differentiate between accuracy and usefulness. Getting more information is sometimes beneficial, but it depends on its price and whether it helps change action. If the main uncertainty is related to management, further clarification of the goal may do little to improve the solution. The state of the world and the state of one's own action model should be taken into account together.

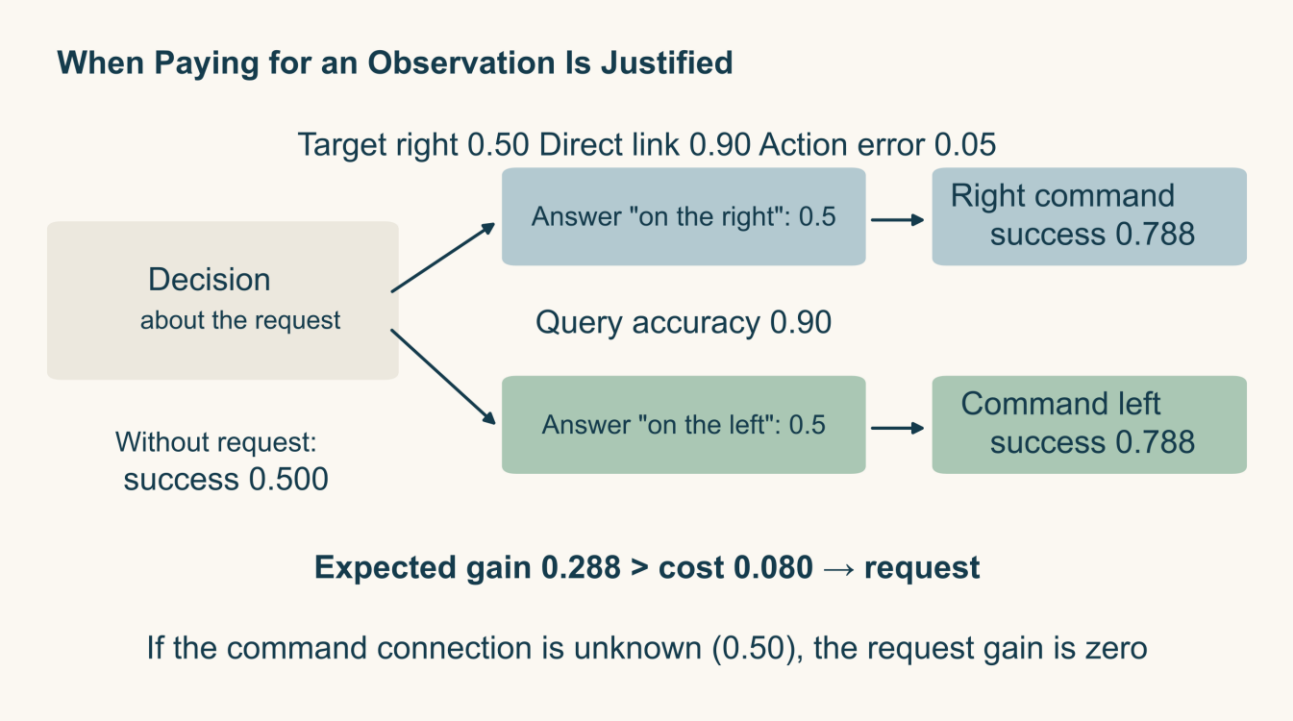

Figure 63. The conditional value of additional observation. Calculation according to agent_lab.py rules for the selected training state at the first missing observation; These are not average launch results. Both possible answers give the expected success of 0.788, so the gain of 0.288 exceeds the price of 0.08. If the probability of direct command communication is 0.5, knowledge of the goal in itself does not improve the choice of action.

Four modes and a fair comparison

Full mode uses all the described mechanisms. No_world_memory only disables the transfer of goal beliefs between steps. Action model learning is preserved. In fixed_action_model, the forward link probability remains equal to the original 0.90, without updating. No_extra_query prohibits the purchase of additional observations. This precise description is important: the name of the mode alone does not explain what calculations were changed.

All modes receive the same potential observations and motor errors for the corresponding scenario. The randomness is divided into independent threads, and the values are pre-prepared. Buying an observation unlocks additional value without shifting subsequent random events. Different commands may produce different effects, but the motor distortion at this step is common.

The attached run runs thirty scenarios with one thousand steps for each mode. These are actual program results, not human or animal data. The request price is 0.08; success brings one, failure brings zero. The utility is obtained after subtracting the query costs.

Mode	Success percentage	Utility per step	Brier
Full	87.94	0.8619	0.1054
Without the memory of the world	85.36	0.7767	0.1202
Fixed Action Model	49.97	0.4841	0.3444
No additional request	77.91	0.7791	0.1589

Brier is the average of the squared difference between the predicted success and the actual outcome, written as zero or one. A lower value means a more accurate probabilistic forecast. It depends on both the calibration and the information content of the forecasts. To call it a pure measure of calibration, self-awareness, or quality of experience would be a mistake.

The overall success of the fixed model of about half hides the sharp difference between the two periods: 87.80 percent before the inversion and 12.13 percent after it. The full mode has the corresponding percentage values 88.06 and 87.83. To measure the rate of retraining, short time windows are needed: the average of the entire second half smooths out the transition process.

There is also a less obvious comparison. The version without memory is more accurate than the version without queries, but it buys significantly more information. As a result, its utility is slightly lower. Therefore, the phrase “this agent is better” requires specifying a criterion. The accuracy of the answer, the cost and the reliability of the forecast may vary in different ways.

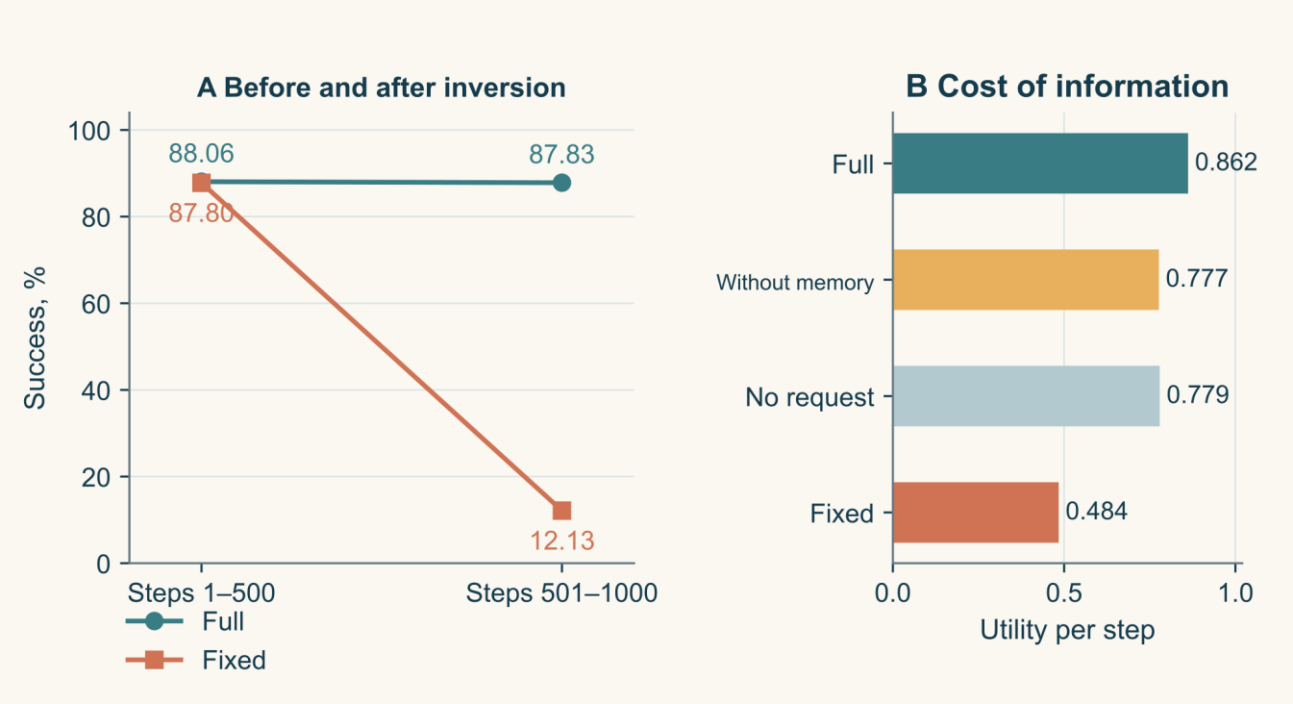

Figure 64. Success and usefulness answer different questions. Results of the agent_lab.py program: 30 general random scenarios for 1000 steps, seed 1729. The average of the two periods is shown on the left, rather than the rate of adaptation within the period. On the right, the cost of additional requests is subtracted from success. Data: lab01_results.json.

How to read one agent step

First, the agent takes into account the possible movement of the target in the stored belief. It then obtains the available noisy signal and updates the estimate. From the uncertainty of the goal and control, it calculates the utility of the commands and the possible request. The prediction of success is written down before the action. After the command, the visible effect updates the control model. The evaluator does not reveal the true goal, inversion schedule and current success to the agent. In this binary task, success along with the effect would reveal the target and change the interface being explored.

What has now been established

We know the order of information and can distinguish prediction from late knowledge. Mode numbers now need to be read along with the asking price and the period before or after the inversion.

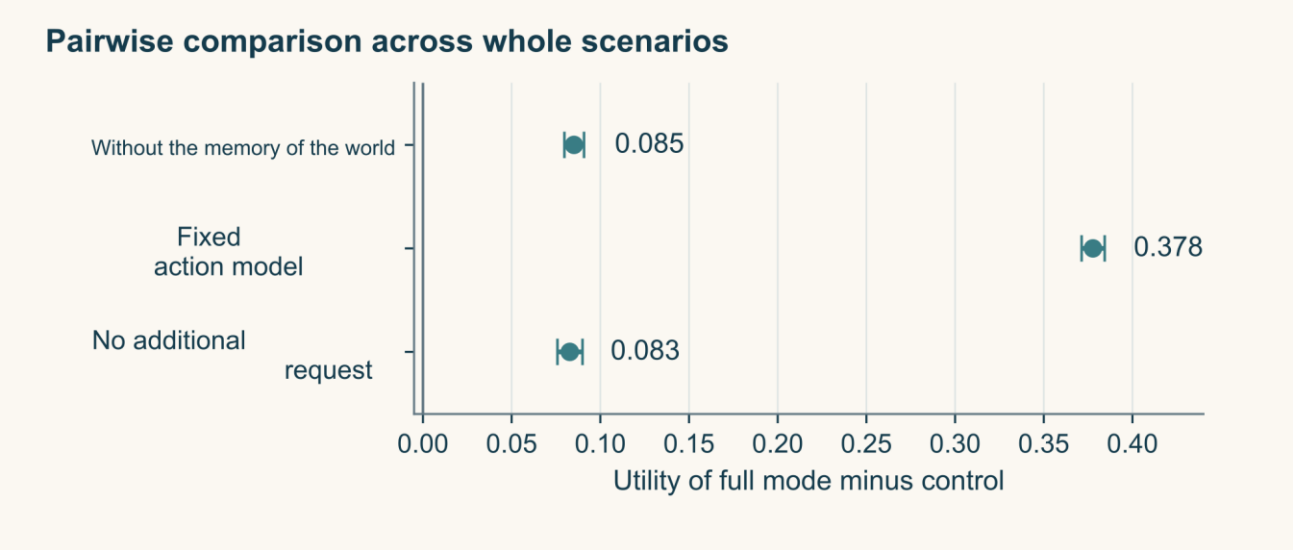

Figure 65. Uncertainty of differences between modes. The result of the analysis of the attached data. Dots are average paired utility differences; intercepts are approximate 95 percentile intervals from 4000 repeated samples of 30 entire scenarios. The unit of repetition is the script. Intervals refer to individual comparisons and do not provide a joint guarantee of 95 interest. Data: paired_analysis.json.

Practice 28 Predict the outcome of environmental changes

Why do it: Reproduce the basic work and explain one result through the design of the mechanism.

What you will need: Probabilistic models of the world and action, Brier score and independent repetition unit. The workshop uses Python 3.10 or later and the standard library.

Time reference: 30–45 minutes. Can be done intermittently; This is not a speed standard.

1. Unpack the practice and run python run_practice.py --self-test. Before the main run, write down the expected behavior of the fixed and learning model after a change in control.
2. Run python run_practice.py --lab 1. Save the created directory, parameters and checksums.
3. In the results table, separately compare the success before and after the switch, the utility per step, and the quality of the prediction. Do not select the best mode without specifying the indicator.
4. Trace one episode through the log: available observation, previous state, forecast, command, visible effect, new state.

What to save: Saved run and short output with mechanism, indicator and condition area.

Hint: The mode without world memory is still learning the connection between the command and the effect. A specific function is disabled, not the entire ability to store the past.

Exploratory continuation: Change exactly one condition by formulating a prediction in advance: request price, sensor reliability, or target variability. Then repeat the comparison on new scenarios.

Analysis and answers - Appendix F, chapter 28, p. 293

Self-test

1. Why does memoryless mode retain learning of the command's connection to the effect?
2. Why is the overall success of a fixed agent misleading without period separation?
3. What would change if you gave the agent the correct answer before recording the belief?

The first agent showed the two beliefs working together. Now let's separately check the presentation device, reprocessing and transmission of content, without confusing them with the success of the entire system.

Chapter 29

Perception and Global Access

Agent success alone did not explain what relationships the representation preserves or why certain treatments help. The three papers in this chapter will break down the problem: the geometry of differences, mixing noisy channels, and delivering content. Each has its own control and its own outcome. After each experiment, we will dwell on what has been established before moving the conclusion to the scientific theory of consciousness.

What exactly should the system represent?

In the previous chapter, the agent distinguished between two sides. Now we are interested in the presentation device: which states the system considers similar, how it changes them during repeated processing, and which modules they become available to. These questions refer to different properties. Convenient geometry does not guarantee accessibility for planning, and wide transfer of information does not explain why one shade is perceived as similar to another.

Let's start with five points on the plane. Each point will represent some object, and the distance between the points will represent the selected difference model. The numbers here are specified by the constructor; they are not measurements of human sensations. Their advantage is transparency: we can specify exactly which relationships are preserved after transformation and which ones change. After this, it will become easier to evaluate more complex models of quality organization.

Complete the second lab. All additional work is in the mechanisms_lab.py file; next to it should be agent_lab.py, which is used by the composite prototype. The result is saved in JSON, suitable for viewing with a regular text editor.

```
python3 mechanisms_lab.py --lab 2 --json geometry.json
```

The program rotates the original points by an angle 0.73 radian and transfers them to the same vector. It then separately stretches one coordinate by half. The distances of each pair before and after the transformation are compared. The slight difference in the last digits when rotated is due to the representation of real numbers in the computer. In a basic run the maximum difference is approximately 0.00000000000000044.

Why invariants are more important than names

An invariant is a property that is preserved under a specified transformation. If the model uses only pairwise Euclidean distances, overall rotation and translation

should not change its similarity predictions. The names of the coordinates may become completely different. The statement “this coordinate is redness” does not stand up to such scrutiny without additional explanation of why this particular recording method was chosen.

Stretching one axis usually changes the distance relationships. This means that it is essential for a model based on these distances. However, the output depends on the selected indicator. A model that uses only the order of objects in the first coordinate can preserve some responses after positive stretching. Before the experiment, you should name what is considered significant: absolute distance, neighborhood order, direction, distance ratio, or the ability to distinguish categories.

Educational geometry does not provide a ready-made map of qualia. Such a map would require data on discrimination or similarity, decisions on the scale of measurement, and testing on new combinations. It is also important to distinguish the distance between stimuli from the relationship between the reported experiences. The same physical step in terms of a parameter does not have to mean the same subjective step. The formulation of the research question precedes the beautiful picture.

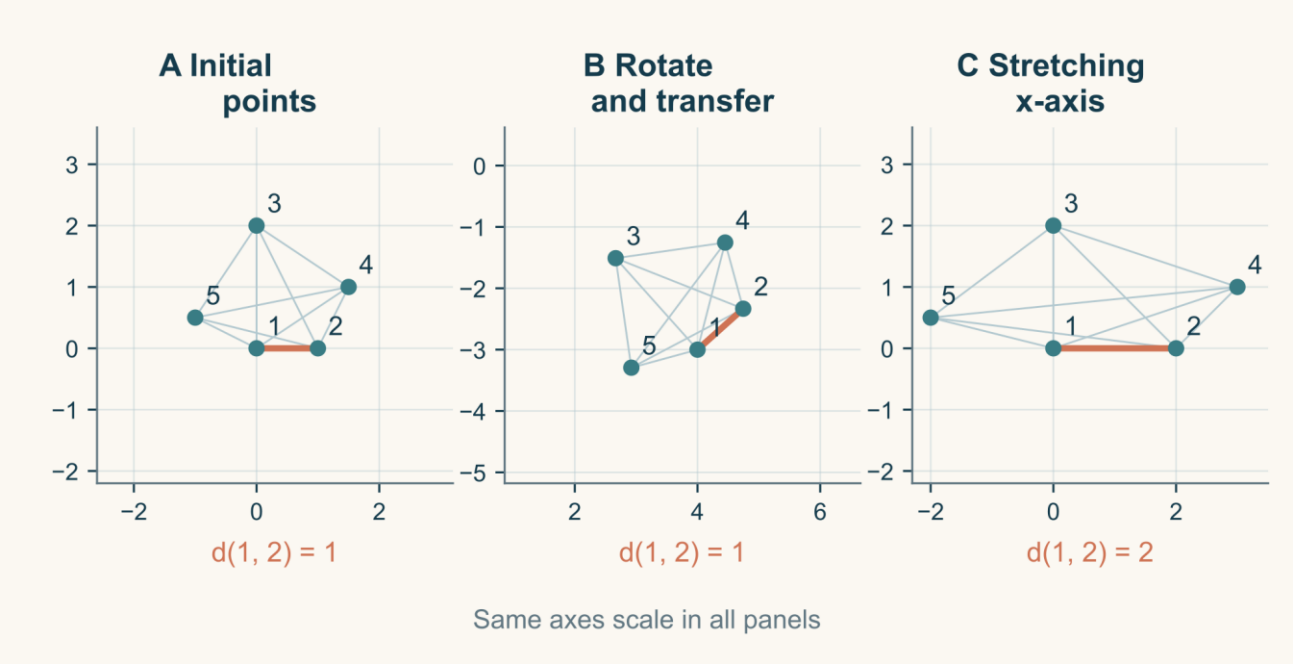

Figure 66. What is preserved when a view is transformed. Educational geometry of the second work. Rotation by 0.73 radian and translation by (4; −3) preserves all ten pairwise distances with machine error; stretching x changes some of the distances. One pair is selected, and the calculation checks all pairs. The numbers represent points, not types of experience. The initial coordinates are specified in mechanisms_lab.py.

Reprocessing with a simple mechanism

In the third work, three channels receive noisy values from a single hidden signal. The true value is equal to minus one or one, noise is added to each observation. One version immediately uses the first channel. The other mixes the channel values four times: half of the new state retains the previous value of a given channel, and a quarter each comes from two neighbors. The first channel is read for the result.

Repeated processing gradually makes the channels similar. This can be understood without a formula. If one channel is accidentally overestimated and the others are closer to the target, the exchange reduces the single deviation. But such an exchange will not correct the common mistake of all channels. It redistributes information already available; a new independent sensor does not appear after each iteration.

```
python3 mechanisms_lab.py --lab 3 --json recurrence.json
```

An important control is four sequentially deployed layers. Each one performs the same operation as one iteration of the recurrent circuit. With the same input, the result coincides exactly. In the first case, the program repeats one step in a loop; in the second, it explicitly records four nested calls. Changing the notation form does not create a new computational property.

Option in 2000 training tests	Mean square error
One channel without mixing	1.892736
Four recurrent steps	0.632419
Four unfolded layers	0.632419
Direct average of three channels	0.632465

The numbers were obtained by the program launched with seed 1729. They are relevant to this task. Direct Average is another useful alternative that uses all inputs without the four mixing steps. Its computational cost is different, so it is not a control with an equal number of operations. The almost equal result shows that there may be a significantly simpler processing method for the selected task.

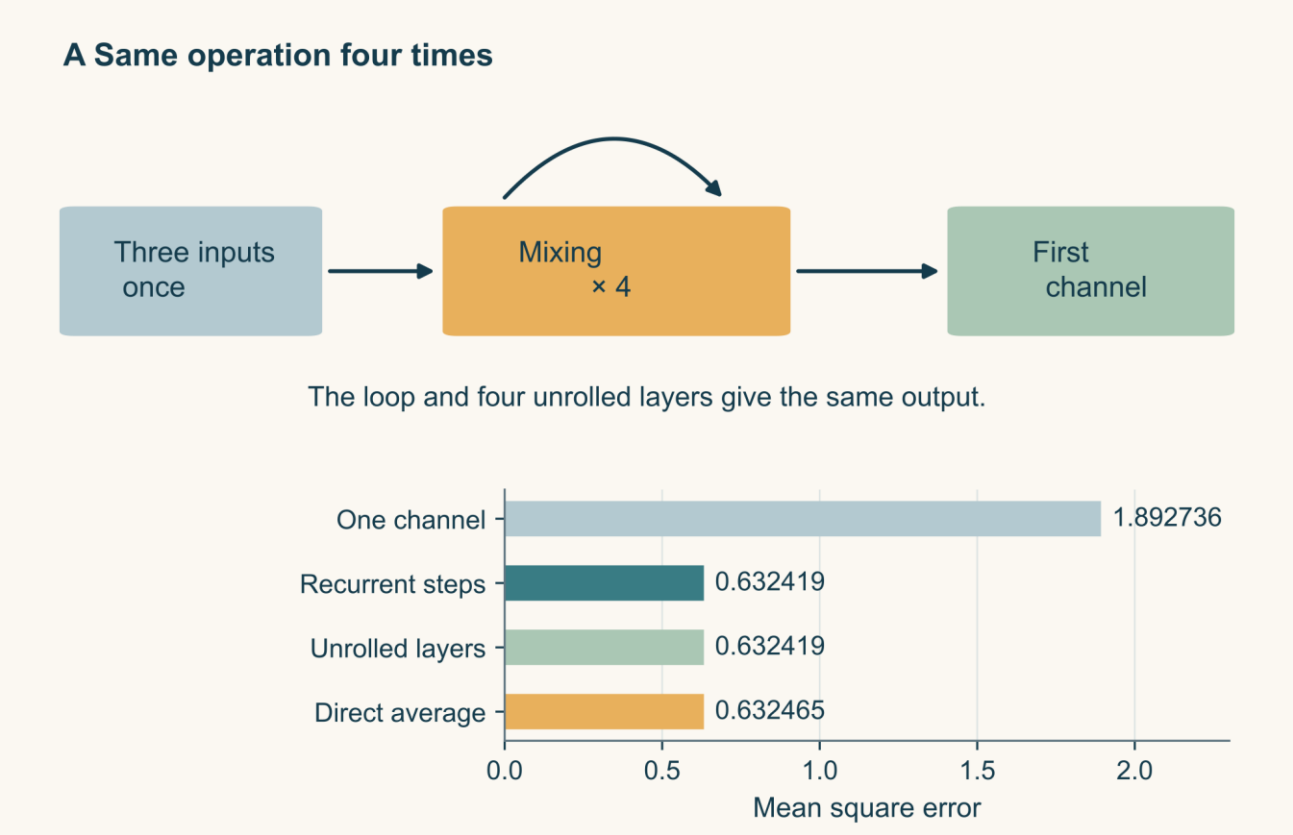

Figure 67. Recurrence and comparable alternatives. Training scheme and results of the third work: 2000 tests, seed 1729. When mixed, a channel retains half its value and receives a quarter each from the other two. The recurrent and unrolled implementations perform the same operations and match exactly. Direct average serves as an additional alternative with different computational costs. Data: mechanisms_results.json, work 3.

Designing a Stronger Test of Recurrence

The final sequence of calculations can often be represented by an expanded diagram. Therefore, the question “is there a loop in the code” is too crude to investigate. It is necessary to clarify which property of the loop is supposed to be essential: reuse of parameters, processing duration, state maintenance, interaction with changing input, or physical mutual causation of elements.

If the hypothesis concerns a computational function, the equality of the final realizations becomes an important constraint. If it refers to a physical organization, the program text does not yet fully describe how the processes are implemented in the device. These differences were introduced in the theoretical part; here we get the opportunity to test at least the computational statement before a stronger conclusion.

To expand the work, you can pre-select a task with different presentation times or with the arrival of new observations between iterations. Then it is necessary to give the competitor the same information and comparable resources. Otherwise, the gain of the recurrent system can be explained by longer observation. A

meaningful experiment occurs when we understand exactly how much architecture difference remains after equalizing the conditions.

Workspace as content transfer

The fourth work uses three messages: color, shape, and distractor. Color and shape take values from zero to two. The correct answer is determined by the sum of color and shape modulo three. The rule is known to the program in advance. We don't need to train a classifier because it learns how to communicate the required information.

The workspace in this model accommodates two messages. The full version selects the color and shape and passes them to the response module. If there is no transmission, it receives an empty representation and uses zero values. With random selection, part of the space is taken up by a distracting message. The additional competitor has direct access to all inputs.

```
python3 mechanisms_lab.py --lab 4 --json workspace.json
```

In three thousand trials, full selection and direct access yield one hundred percent correct answers. Without transfer, the success rate is 34.40 percent, with random selection - 55.83 percent. These results are expected from the task device. The relevant channels are named by the constructor and the response rule is specified. The program did not independently discover what was considered important.

This experiment confirms the operation of the transmission channel. It also shows that this workspace is not the only way to solve the problem: direct access gives the same result. If we want to explore the benefit of limited exchange between specialized modules, a more complex set of tasks and meaningful comparisons with other modes of coordination will be required.

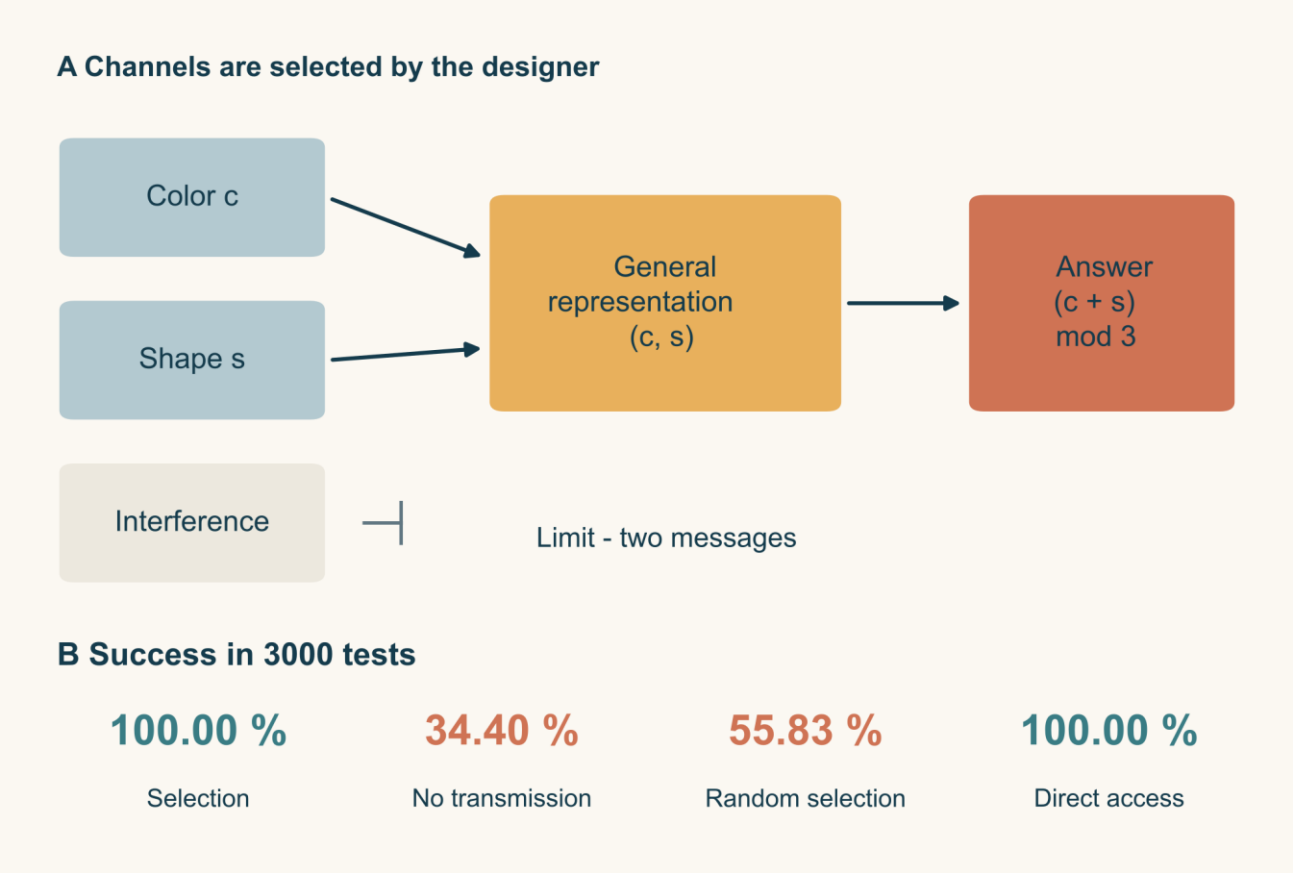

Figure 68. Content delivery and direct access control. Training scheme and results of the fourth work. Color and shape have the values 0, 1 or 2; the response rule is predetermined. The full version conveys the two required messages. A competitor with direct access solves the problem just as successfully. Here the delivery of information to a given rule is checked. Data: mechanisms_results.json, work 4.

From transfer to flexible use

Imagine the continuation of the experience: after perception, a new instruction comes. In one case you need to compare the color, in another you need to use the shape, in the third you need to remember the object for the next step. The system should preserve content that is useful to different consumers. This brings up the issue of flexibility of use rather than just delivering two numbers to a fixed rule.

However, new instructions must also be given to the competitor. If only the workspace receives task information, the benefit will be trivial. A useful design prefixes the amount of memory, the number of messages to be transmitted, the processing time, and the set of possible instructions. You can test transfer to new combinations without changing the selection rule after each unsuccessful result.

The next step is to disable content distribution itself, while maintaining local processing. Then compare this intervention with turning off an equally costly but insignificant channel. Such controls help figure out which change is important.

Degradation when a large part of information is removed in itself says little about the special organization of consciousness.

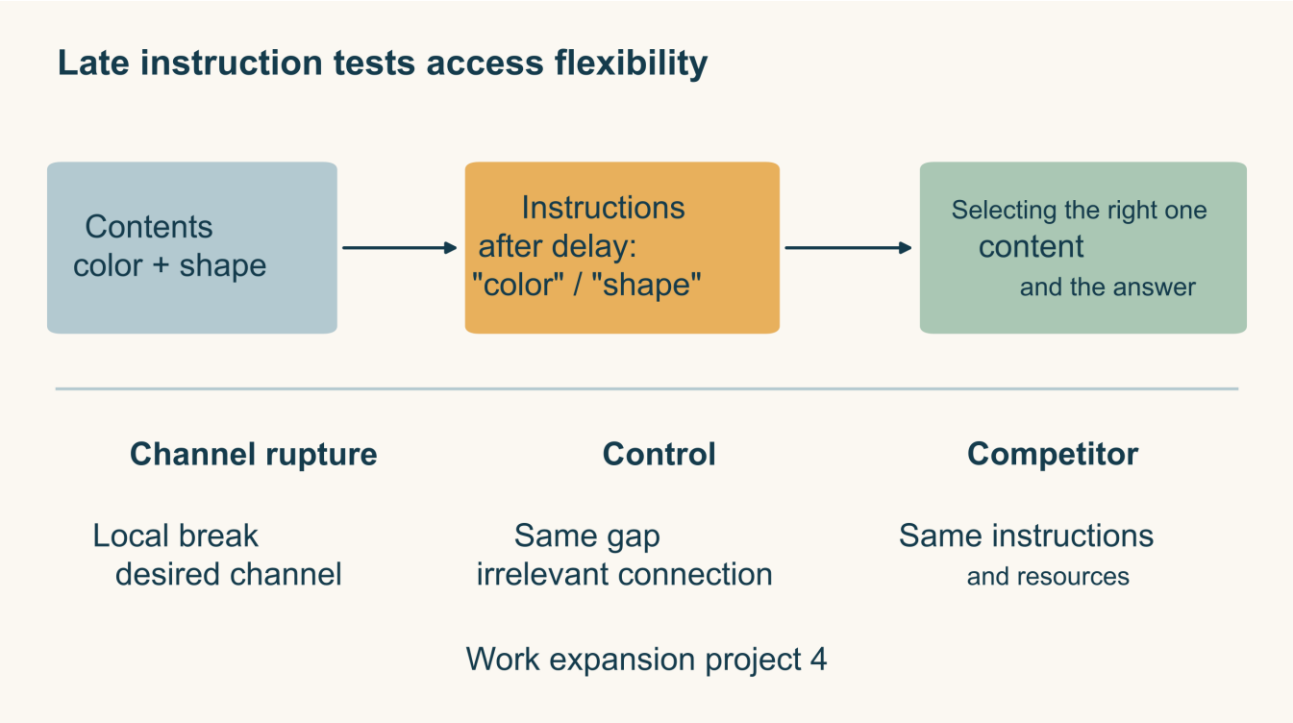

Figure 69. Access check after late instruction. In the proposed experiment, the system recognizes the desired task after presenting the content. The selective intervention is compared with an irrelevant gap and a competitor with the same resources; this design tests more than the delivery of two numbers in a fixed task.

Why Equal Outputs Limit Architectural Inferences

The four recurrent mixing steps can be written as four sequentially unwrapped layers with the same rule. If the input is the same, the calculated result will be the same. This means that a specific gain over one noisy channel cannot be explained by the mere fact of the presence of a loop in the source code. It is necessary to clarify what is assumed to be significant: the number of available inputs, time, reuse of parameters, a changing signal, or the physical organization of the execution. A useful control preserves the remaining differences and checks the selected ones.

What has now been established

Geometry, processing and transmission describe different properties. One lucky number does not replace three checks. Now let's choose one work and bring its argument to the end.

Practice 29 Make two discriminating experiments

Why do it: Test one architectural statement against a strong control case.

What you will need: Invariant from chapter 2; feedback and comparable resources; local state and shared access. Work runs 2–4 are independent and do not require changing all mechanisms at once.

Time guide: 30–45 minutes for one job. Can be done intermittently; This is not a speed standard.

1. To work with 2, select the invariance indicator in advance: a complete list of pairwise distances. Run python mechanisms_lab.py --lab 2 --json geometry.json and compare rotation with stretch.
2. To work 3, first predict the difference of one channel, blend, unwrapped layers and direct average. Perform a separate run --lab 3.
3. To work 4, write down which messages actually reach the response rule. Compare full selection, no transfer, random selection and direct access via --lab 4.
4. Choose one of the three experiments and state its immediate result, a strong competitor, and the next differentiating question.

What to save: One saved experiment with verified controls; the rest of the work can be completed in separate sessions.

Hint: The coincidence of one pair of distances does not establish the invariance of all relations. Winning over a version without the required inputs does not establish the uniqueness of the architecture.

Research continuation: For mixing, add a common noise component to all channels and save the independent one separately. For transfer, design a late instruction with equal inputs and competitor resources.

Analysis and answers - Appendix F, chapter 29, p. 294

Self-test

1. What does the persistence of distances show after rotating the view?
2. Why does the coincidence of the recurrent and expanded versions not solve the question of all theories of recurrence?
3. Which conclusion is supported by the equal success of workspace and direct access in the fourth work?

Individual functions received their own controls. The next chapter will add source history, internal access, and strategy revision, and then test their implementation together.

Chapter 30

Self-model and reflection based on memory

This chapter contains several different studies. Go through them sequentially: first the source history, then the internal access, then the change in strategy, and only after that the composite agent. The failure of one comparison does not invalidate the results of another; the success of a single module does not automatically transfer to the entire system. For each step we will save the device, intervention, result and the exact limits of inference.

From own action to source history

When an agent learns to associate a command with an effect, it has a useful model of its effect on the environment. However, the experienced authorship of an action, bodily belonging, and explicit reasoning about oneself remain distinct objects of explanation. In this chapter we will build several small mechanisms and find out what allows each result to be asserted. All experiments are performed in code; the transition to the experience hypothesis is discussed based on the requirements of chapter 26.

The first mechanism continues the example from chapter 25. Let's imagine two observable sources A and B. For each event it is known whether the current sign supports the hypothesis of a common cause of signals. In the training example, source labels are available to the model and are recognized accurately. They are not hidden marks of the body's true identity. This idealization allows us to first test the mathematics and then separately examine source recognition errors.

The M0 model has no belief transfer: each event begins with a probability 0.5. M1 stores one belief for the entire thread. M2 stores a separate belief for each observed source. The general identification tracker M1I does the same for external objects. In the minimal implementation, M2 and M1I are intentionally mathematically the same. This is a strong control: the memory name should not artificially create an advantage.

Executable example of shared and separate memory

A positive sign has a probability of 0.75 for a common cause and 0.25 for separate causes. Before the update, the previous belief is one-fifth closer to 0.5. For M2, this happens the next time a given source is accessed, and not at each world step. A

different model of forgetting would yield different meanings, so the rule must be written explicitly.

```
python3 mechanisms_lab.py --lab 5 --json source.json
```

In the source_history_example section of the result file there is a sequence A is positive, B is negative, A is positive, B is positive. The program reproduces values that can be checked manually. The table shows the updated estimate of the active source; the inactive entry M2 is stored separately.

Event	M0	M1	M2	M1I
A positive	0.7500	0.7500	0.7500	0.7500
B negative	0.2500	0.4375	0.2500	0.2500
A positive	0.7500	0.7105	0.8750	0.8750
B is positive	0.7500	0.8581	0.5625	0.5625

At the third event, M2 returns to the positive history A, while M1 also suffers the negative experience B. At the fourth event, the opposite side appears: M2 returns to its own negative history B. Separate memory does not always have to give a higher probability. Its distinctive property is the dependence of the update on the history of the corresponding source.

These values are consequences of a given algorithm. They do not yet show how a person’s experienced belonging changes. To relate models to such data, one must select a measurement, estimate parameters on one set of observations, and test predictions on another. If M1I continues to give the same answers, it is the special mechanism of self-reference that has not been identified. It is necessary to find an additional theoretical difference or to recognize the general nature of the explanation.



Figure 70. Whose history is used when updating. Results of the fifth work. The dots show the assessment of the active source after the event; the lines connect sequential events and do not represent a continuous change in the record of one source. M2 and the general tracker M1 are computationally the same. Marks A and B are observable and unmistakable here. Data: mechanisms_results.json, source_history_example.

How to find the source of your own influence

The second part of the fifth work solves another problem. The agent sends binary commands and sees two streams of consequences. One is related to commands, the other is random. The associated thread can be two steps late. The model must determine the channel from the observed history, without access to the hidden index of the managed source.

One version compares only simultaneous values. The other considers delays from zero to three and selects the best match. The general tracker receives the same information and uses the same calculation. Additional control randomly shuffles command history. First, a channel and delay are selected based on one hundred events, then the match is checked using sixty new events. In this way we separate the choice of explanation from subsequent testing.

With zero latency in the basic three hundred repetitions, both basic methods correctly determine the channel. With a delay of two, simultaneous comparison gives 49.67 percent of correct choices, history and the general tracker - one hundred percent each. On new events, the connection chosen by the historical

model is reproduced in approximately 87.7 percent of cases, which is consistent with the inherent noise. The shuffled history produces results around the random level.

Many observations are combined to select a channel. Therefore, an error-free channel determination is compatible with prediction errors of individual future effects.

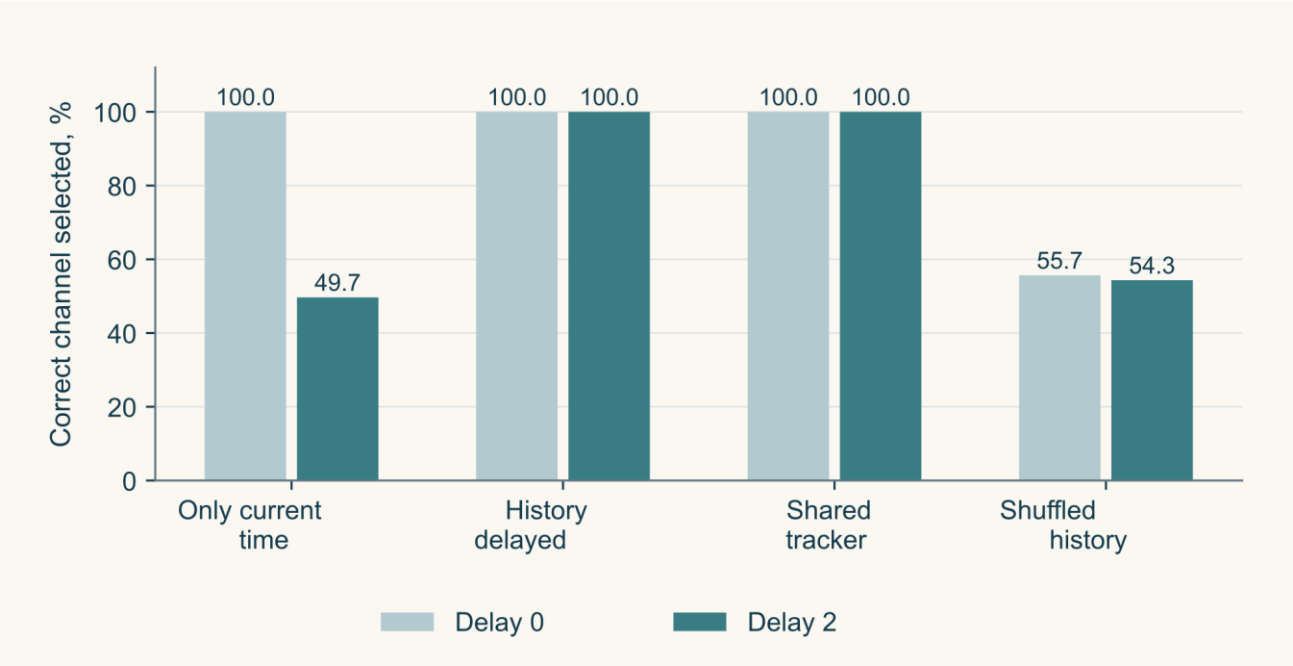

Figure 71. History helps to recognize a delayed source. Results of the fifth work: for 300 repetitions at delays 0 and 2 step. The bars show the correctness of channel selection across many events, rather than the proportion of correct predictions of individual effects. The history and general tracker use the same algorithm. Channel and delay are selected at 100 events; compliance is additionally checked for 60 new events. Data: mechanisms_results.json, conditions_by_delay.

Internal access with causal verification

The sixth work creates a minimum standard for internal monitoring. There is a hidden binary feature in the system. Public action does not depend on it. The internal sensor receives the sign directly and reports the value; the outside observer only sees the public side and uses the basic response. Another version has a disabled internal channel.

Nothing changes in the control condition. During sham intervention, an extraneous bit is changed. With targeted intervention, it is the internal state that is being assessed that is reversed. The report is checked against the condition after the intervention. All conditions receive comparable initial events, so one can ask whether the message follows the desired cause.

```
python3 mechanisms_lab.py --lab 6 --json internal_access.json
```

The internal sensor responded correctly in all two thousand trials of each condition, while the external observer and the disabled version responded correctly in about half. The result is as expected: privileged access has been explicitly designed. We have obtained a proof-of-concept with which more complex systems can be compared. When studying a language model or a self-trained agent, the presence of such a channel will still need to be established.

The structure of the test remains the same as the complexity of the problem increases: what variable is available, when is it available, and what alternative information could trigger the response? The coincidence of the report with the experimenter's expectation allows for different explanations.

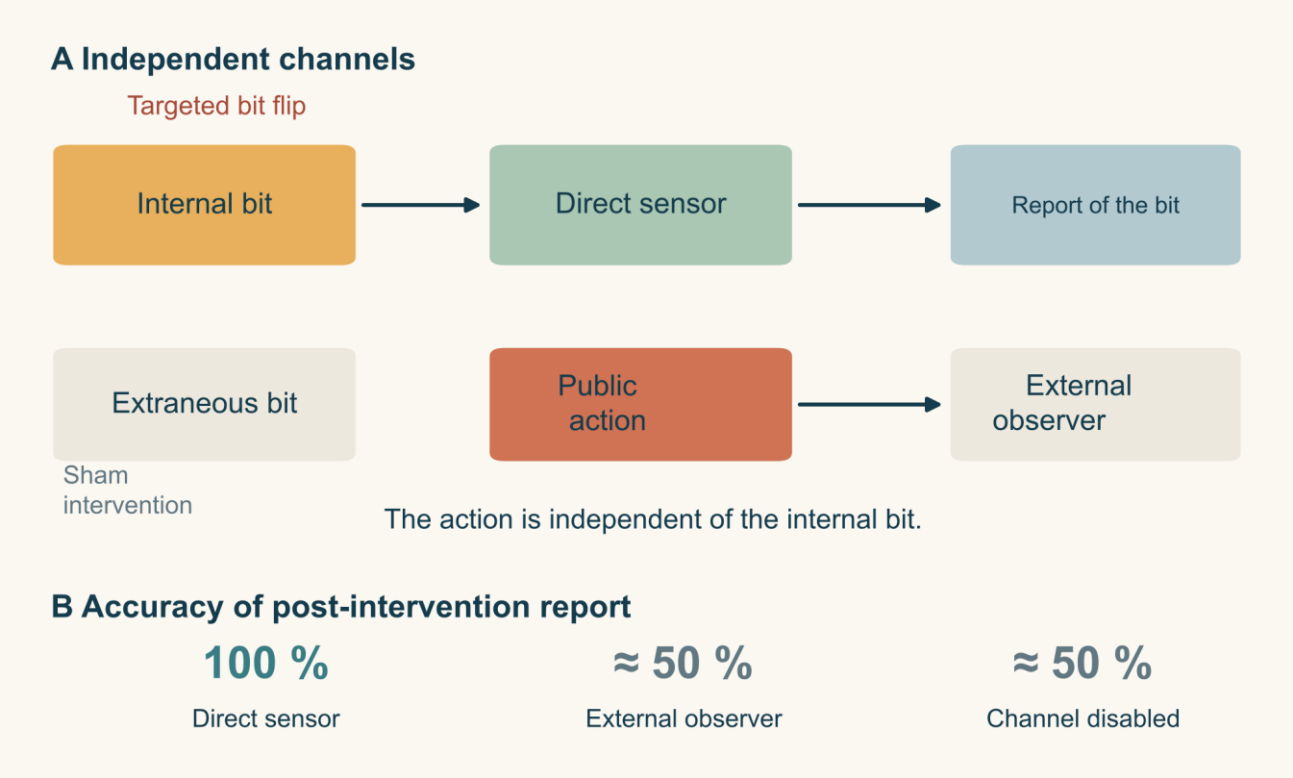

Figure 72. Internal access is verified by intervention. Training scheme and results of the sixth work: according to 2000 tests in the initial condition, with a fictitious impact and when flipping the target bit. In all three conditions, the direct sensor gives 100 percent; two controls - about 50 percent. The assessment refers to the post-intervention status. The access channel is specified by the designer; the diagram does not depict the detected introspection of complex AI. Data: mechanisms_results.json, work 6.

Revision as a reason for new action

In the seventh job, the agent chooses between two predefined classification strategies. Each compares the observed number with its own threshold. The first is true before the environment changes, the second - after. The agent does not receive a rule change flag. The correct answer is revealed only after recording the current prediction and can be used in the next steps.

The adaptive version stores the thirty most recently marked examples. Before each new answer, it compares the errors of both strategies on past data and chooses the best one. The fixed version retains the first strategy. The narrative_only version could talk about revision, but its actions remain fixed. In the code, the language story is not generated: the name denotes control of the causal influence of the intended explanation.

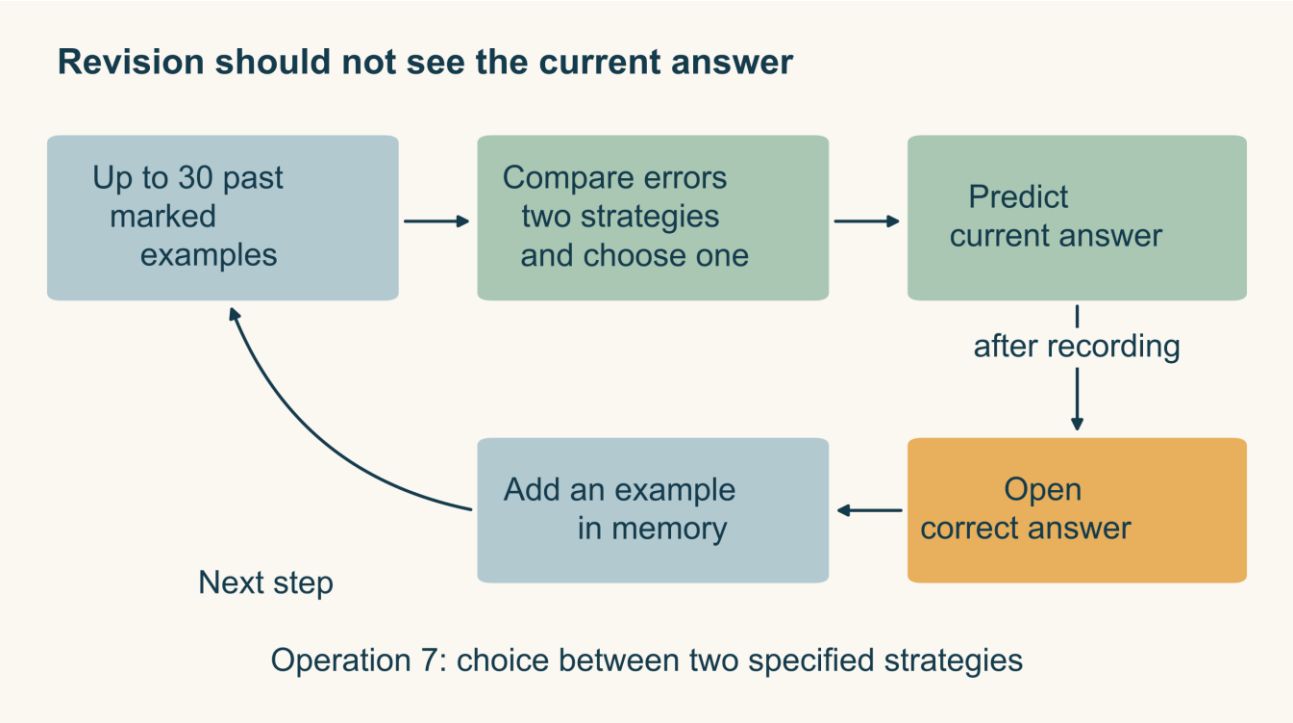

Figure 73. Causal order of functional revision. The order of operations follows revision_lab(): the current correct answer is only available after the prediction. No more than thirty past examples come into memory; adaptation chooses between two predetermined rules and can only influence subsequent decisions.

```
python3 mechanisms_lab.py --lab 7 --json revision.json
```

In the baseline experiment of 1200 steps, the adaptive version changes the strategy once. Before the change, all modes respond correctly; after it, adaptive gives 99 percent of success, fixed and control without changing actions - 69.67 percent. The JSON stores a section of log around the switch, which shows which strategy was chosen before the current response was received.

This is a small model of functional revision. We have specified the candidate rules and the method for comparing them. More developed reflection will require the formation of alternatives, evaluation of initial assumptions, transfer of conclusions and control of one’s own errors at different levels. But already the initial model allows us to test the key causal relationship: the revision should change subsequent decisions.

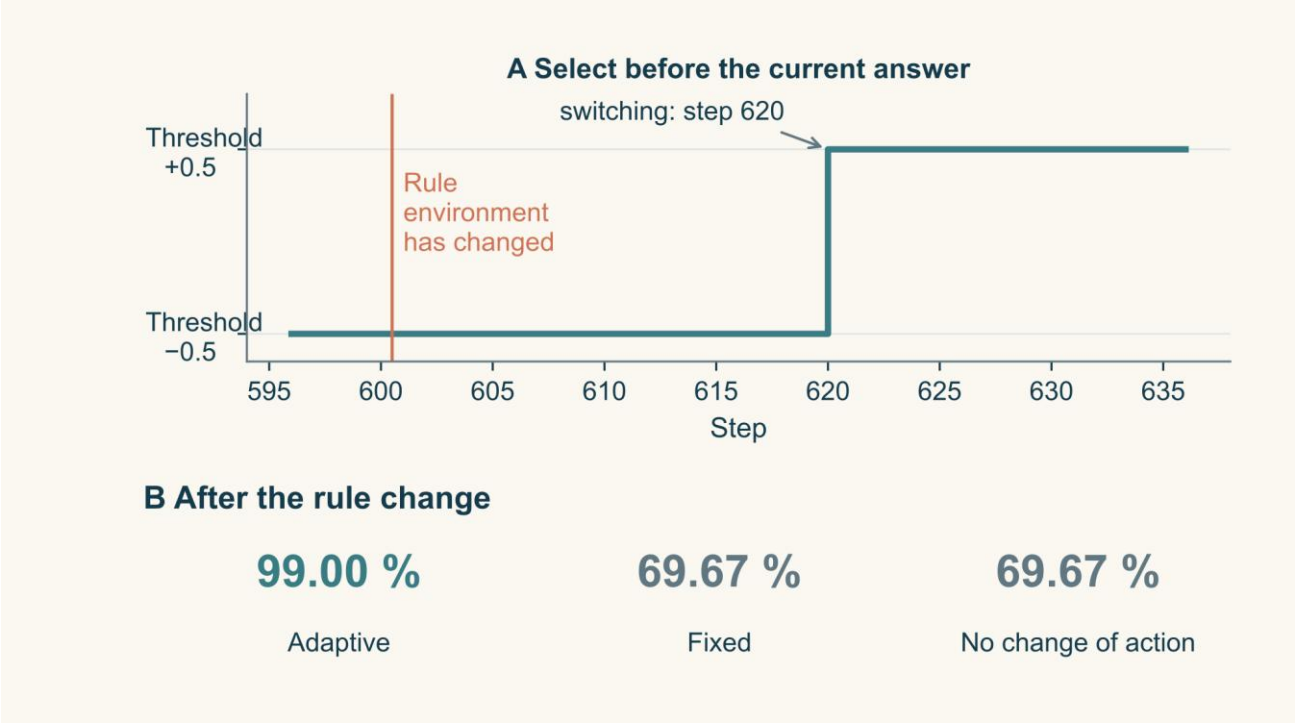

Figure 74. The revision becomes the reason for a new decision. Results of the seventh work, 1200 steps, seed 1729. The saved section of the log is shown: the strategy changes at step 620 after errors in past responses. The correct label for the current case is revealed only after the prediction. Below - success at steps 600 after changing the environment. The narrative_only control saves actions; The program does not generate a text story. Data: mechanisms_results.json, work 7.

Composite Candidate Assembly

The eighth work combines several mechanisms in one prototype. It uses a probabilistic world memory, a self-influence model, a shared two-belief representation, a choice module, a predictive success monitor, and a learning rate change rule. The medium is taken from the first work. So we have a functioning system with several interacting parts.

The overall view publishes the probability of the target on the right and the probability of direct control. The action module and monitor read the same version of this information. A history of unexpected effects may enable more rapid revision of the management model. The candidate is not given the correct goal or schedule for changing the environment. The researcher saves the log and calculates the metrics separately.

```
python3 mechanisms_lab.py --lab 8 --json candidate.json
```

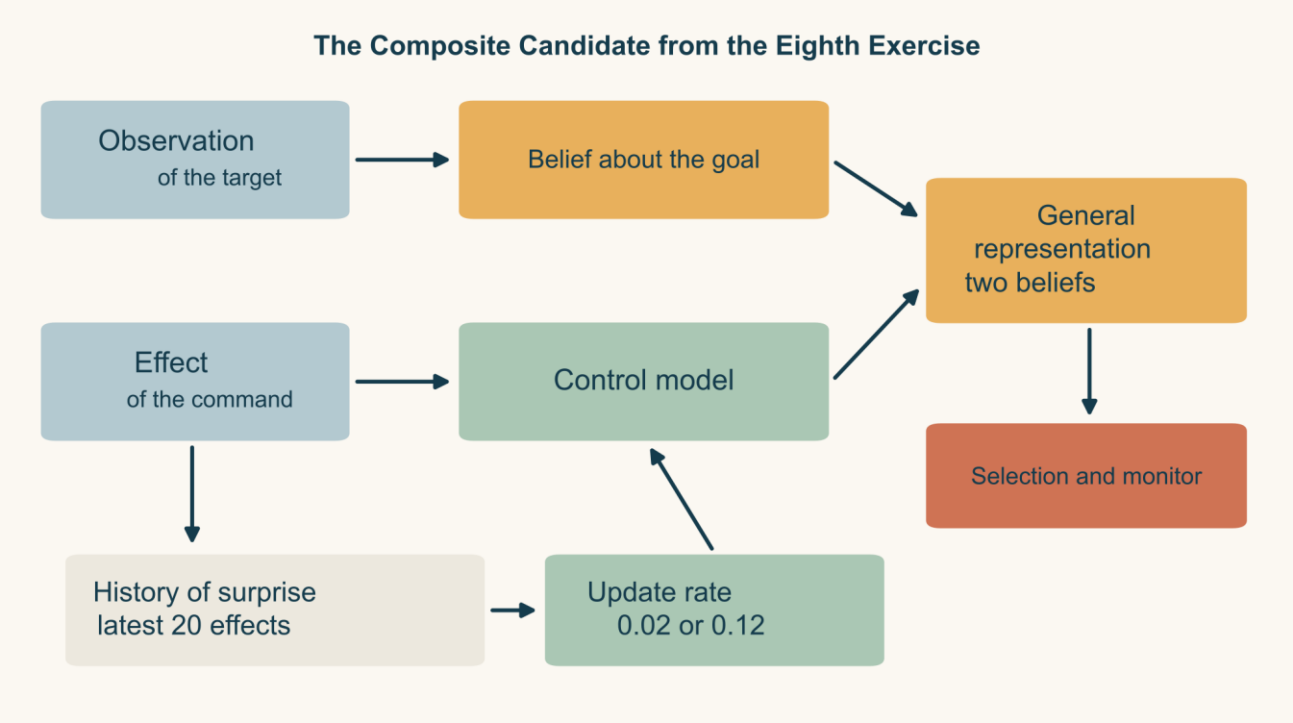

Figure 75. Assembling several mechanisms into one system. Study diagram for the eighth work. Two beliefs are published for action selection and success prediction. The average surprise of the latest 20 effects may switch the control model revision parameter from 0.02 to 0.12 when the threshold 0.35 is exceeded. This conditional rule is designed explicitly. Arrows indicate information transfer and parameter changes; they do not establish experience.

In a thousand-step scenario, the full version gives 83.80 success percentage and 0.81808 utility per step. Direct reading of beliefs gives the same result. The no_broadcast mode simultaneously disables the transmission of content and additional observations; its success rate is 51.90 percent. Therefore, the comparison does not highlight the contribution of one channel. This is a separate scenario, not the average of thirty runs of chapter 28.

The fourth mode is especially useful, where the rate of revision is fixed. In this launch it coincides with the full mode. The log shows the reason: the threshold for enabling accelerated learning was never reached. The module is present in the program, but its contribution was not evident in this experiment. The parameters were not changed after viewing the result to gain an advantage.

This gives rise to the following project: select in advance the conditions where the mechanism should turn on, and test them on new scenarios. The frequency of control changes or the rule for assessing surprise can be examined. If parameters are selected based on data already seen, this stage should be called search. For the verification phase, the conditions are fixed until a new launch.

How does the presence of a module differ from its contribution?

In a composite candidate, the expedited review rule is enabled only if the surprise reaches the threshold. In the base scenario this never happened. Therefore, matching the full version and the fixed tempo version does not compare the effect of the two tempos at the moment of the actual switch. First you need to check the inclusion log. Then, under new pre-selected conditions, situations where the rule should work can be explored. Selecting a successful threshold based on the initial result will be a search step and will require a separate check.

What has now been established

Now each function has its own question: is the history used, is the internal state available, is the revision changing the new action, and is the component required in the composite agent.

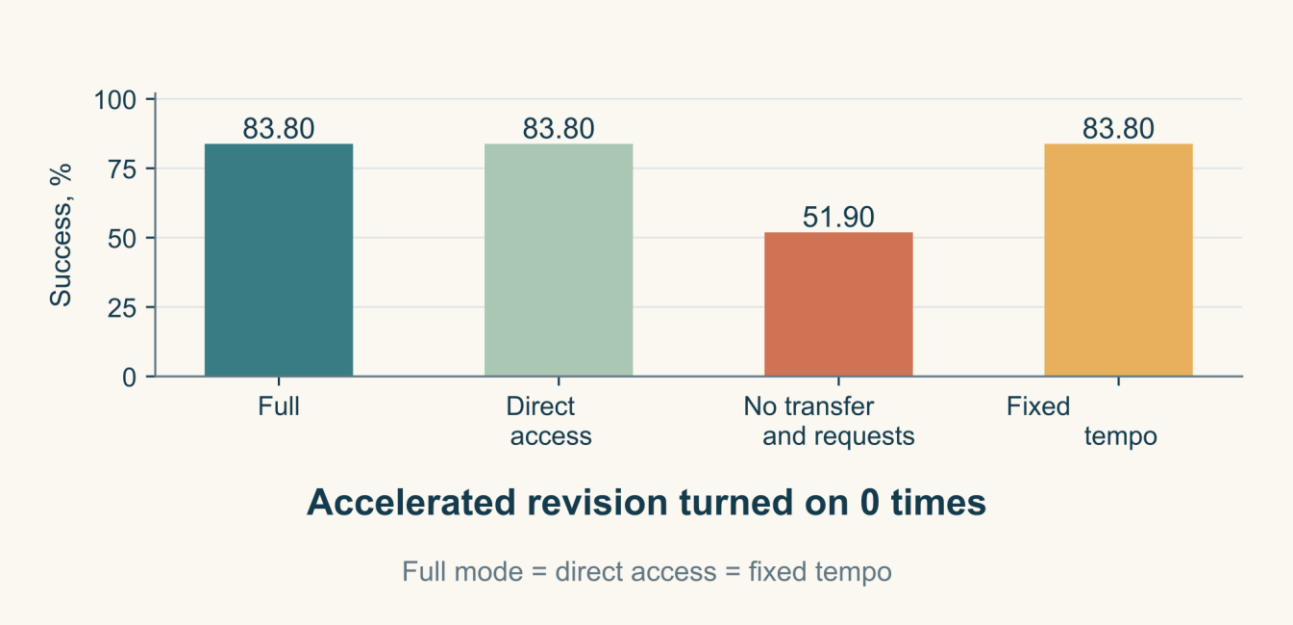

Figure 76. The presence of a module does not yet indicate its contribution. Results of the eighth work: one scenario from 1000 steps, seed 1729. The three regimes coincided on all saved indicators: the accelerated review condition never worked. No-transfer mode also disables additional requests, so this is a compound intervention. To verify an individual contribution, a new predefined experience is needed. Data: mechanisms_results.json, work 8.

Practice 30 Protect the conclusion about your own system

Why do it: Prepare a causal argument for one selected work instead of summing up all successful demonstrations.

What you will need: Separate memory from chapter 25; causal access check; preliminary forecast and feedback. Works 5–8 can be performed on different days.

Time guide: 35–50 minutes for one job. Can be done intermittently; This is not a speed standard.

1. Select job 5, 6, 7 or 8 and execute it separately by running mechanisms_lab.py with the appropriate --lab.
2. Before examining the result, indicate the available inputs, the function being changed, and the main indicator.
3. For operation of 5 compare the general tracker; for 6, check target and dummy intervention; for 7, save the forecast until feedback; for 8, check the number of revision inclusions.
4. Write four sentences: what is realized, what is causally verified, what remains undifferentiated, and what next experiment is needed.

What to keep: One testable report with a log, a control case, and a separate conclusion about the phenomenal hypothesis.

Hint: A comparison that simultaneously disables transmission and additional requests does not highlight the net contribution of a single transmission.

Research continuation: Formulate new conditions for activating the selected component before launch. Keep the previous non-discriminating result and compare the strong competitor in new scenarios.

Analysis and answers - Appendix F, chapter 30, p. 294

Self-test

1. Why does the equality M2 and M1I represent a meaningful result?
2. What determines one hundred percent correct responses of a designed internal sensor?
3. Can we consider the benefits of accelerated revision as shown in the eighth paper?

The digital model makes all states available to the developer. When moving to living tissue, some of these possibilities disappear, and the role of external equipment increases. Therefore, the next chapter will separately prepare work with real data.

Chapter 31

Biological and hybrid implementations

Moving to a biological implementation changes the available measurements and the boundaries of the system. Some learning may take place in an external controller, and many cycles may belong to the same sample. We therefore first need to learn to read a real dataset and check the units of observation. A complete teaching route through publicly available fields follows below; an independent search for a measure of interaction is a further step after reproducing the example.

What changes when living tissue appears

The software model allows you to precisely specify the state and transition rule. The biological system provides other research material: cells develop, interact and change their activity patterns. The ability to use these properties for calculation does not mean that any living entity has experience. But it is usually not enough to reduce a biological project to running a ready-made algorithm: it is necessary to establish what functions the tissue itself performs and what external equipment provides.

We will distinguish between a culture of neurons, an organoid and a whole organism. For our analysis, the structure of the object being studied is essential, and not the impression of the word “brain”. A hybrid is a system where the biological component is causally related to the digital transformation of inputs, activity reading, and task. The boundary of the intended bearer of experience may not coincide with the boundary of the cup or computer process. It will need to be justified.



Figure 77. Biological component and measuring equipment. Subject illustration of a conventional research workplace. The cell model in the dish, the equipment and the computer are shown as different parts of the system. Their actual connections and contribution to the task must be described in a specific project; the figure does not reproduce the setup of any of the experiments discussed.

Three research examples

In DishBrain, neural cultures of human and mouse origin were linked through electrical recording and stimulation to a simplified game of Pong. The authors reported improved gaming performance with certain closed-loop feedback. These are cultures, not organoids. The word sentience in the title of the article is associated with the author’s operational interpretation; the measured outcome does not establish the subjective experience. [44]

Brainoware used the organoid as a dynamic computational reservoir. The work includes speech signal recognition and nonlinear dynamics prediction. To understand such a device, it is necessary to separately consider the transformation of the input, the response of the biological network, and the procedure for reading the result. The success of the overall system does not in itself indicate which component provided a particular improvement. [45]

Robbins and colleagues examined cartpole control of cortical organoids. In most samples, training cues selected by the reinforcement algorithm provided an advantage over random cues or no training. The improvement did not persist after a forty-five minute rest. This result distinguishes between adaptation during the

lesson and the maintenance of the achieved effect after a break; it does not mean the absence of any forms of memory in organoids. [46]

Complete a chain of levels

Suppose a biologist proposes to change a molecular process to enhance a certain form of plasticity. To connect this proposal with our hypothesis of experience, we will have to go through several transitions. First you need to show a change in the molecular process. Then there is its connection with the properties of cells. After this, the network interactions change. Next are the consequences for the task. Finally, separately substantiate the relationship of the resulting organization to the experience.

Each arrow may be ambiguous. Altered network activity may be explained by cell composition, maturation, general excitability, measurement conditions, or reorganization of connections. If the author moves directly from molecular intervention to “consciousness enhancement,” intermediate explanations remain untested. The research project must name a measurement or comparison for each significant transition.

In this book we do not choose a molecular target as a guaranteed way to create a subject. We learn to identify which change tests a particular hypothesis. For example, if mutual dynamics are assumed to be significant, it is of interest to change certain interactions while maintaining other capabilities of the system. A simple increase in overall activity does not provide specific support for such a hypothesis.

Two mechanisms for the same improvement

Let's build a training example. Digital Model A learns to change the connections between several components, so that the same input leads to the desired answer more and more accurately. Model B retains the connections, but changes the overall gain. On a simple sequence, both can improve the task score. This is a possible design for testing the analysis, rather than a description of the results of said biological work.

If we measure only the outcome of success, the mechanisms remain indistinguishable. Let's add new combinations of inputs. Changes in specific connections may predict selective transfer between combinations; overall gain is a different profile of change. The direction of difference must be inferred from the selected models. If the rules have not yet been set, the words “selectivity” and “plasticity” by themselves do not create predictions.

You can then search public data for the metrics needed to make this comparison. There may be insufficient data: only general results are recorded, the required inputs are not stored, or control is missing. This is a meaningful result of the preparation. It shows what information the next biology project should provide.

A Guided Route through an Open Dataset

For independent work, we use the specific file aggregated_experiment_data.pkl from version v2 of the Robbins repository. Checksum MD5 file af99f1f859136ca4a6dad47a1aa7c86d coincides with the published one. The repository description reports nineteen organoids and four hundred and thirty-two cycles. The checked file contains 38 top-level experiment records, 426 run records, and 16 various metadata.organoid values. These units must be kept separately; the number of field values in a given file does not override the selection of the entire article. [47]

The set boundary is part of the method. The chip field retains the original designation, organoid is intended for a purified grouping, batch indicates a batch. One experiment record can have multiple runs, and one organoid can have multiple experiments. Therefore, randomly rearranging all rows does not create independent patterns.

In the attached table, the characteristics are obtained using one explicitly defined procedure. The frequency is the average of the final firing_rates_hz. The intensity of temporal relations is calculated as the average of the absolute elements of the upper triangle sttc_first_order without the diagonal and non-finite values. Using the modulus does not distinguish between positive and negative relationships, so a different hypothesis would require a different measure. STTC should not be equated with causation without further argument.

The training comparison target is the median of the top_decile among the available rl runs within the experiment record. The documentation defines top_decile as the average reward of the top ten percent of samples. We do not replace this value with overall accuracy, overall learning, or retention of effects after rest. A different outcome would require a new description of the problem.

Of the 38 rows, 29 are usable: both features, the target measure, and the organoid identifier are available for each. They cover 15 organoid identifiers and 7 batches. The remaining rows are retained together with their reasons for exclusion. A missing measurement is not replaced by zero, and exclusion does not depend on whether it improves the desired result.

The basic model uses the logarithm of one plus the average discharge rate. Advanced additionally uses the selected STTC indicator. Transformation and

scaling are evaluated only on the training part. External verification lays aside the whole organoid one by one; a separate option postpones the entire batch. The result applies to these procedures and a limited set.

This analysis demonstrates a reproducible relationship between tabular characteristics and a given indicator. Prospective prediction requires independent confirmation of the temporal order of measurements; causal inference requires intervention design. Adding a feature does not necessarily improve the error. We keep the negative result equally with the positive and do not choose a convenient grouping after viewing.

One line of data and one clearly defined indicator

In the bio_features.csv file, one line corresponds to an experiment record, not one neuron or one organoid. The mean_firing_rate_hz field averages the final values of the original firing rate. The indicator mean_absolute_sttc15 is equal to the average of the absolute finite elements of the upper triangle of the sttc_first_order matrix without a diagonal. It characterizes the severity of temporal relations of activity in the selected window; is not a proven physical relationship or measure of experience.

For the line 20217_exp1 we get approximately 5.35918 Hz, 0.0172533 for this indicator and 108.95 for the top_decile median in rl runs. According to the documentation, top_decile is the average reward of the top ten percent of samples of an individual run. Our additional aggregation between runs is specified explicitly. It serves the learning task of associations rather than assessing the causal effect of learning.

What has now been established

We know the origin of the string, the meaning of the features, and the bound of the result. It is now possible to test whether a selected feature adds statistical information to a given procedure without turning the observed association into a causal explanation of the tissue.

Reproducible analysis result

The attached calculation uses 29 of valid experimental records. A linear model with a constant compares log(1 + frequency) with the same trait and the mean absolute STTC. Features are standardized only in the training part; the ridge regularization coefficient is 0.000001. The main result is the mean square error with equal weight for each delayed group. The smaller it is, the more accurate the forecast in a given procedure.

Deferred group	Number of groups	Frequency	Frequency and STTC
Organoids	15	1460.17	1581.03
Parties	7	1658.10	1700.39

Adding this STTC metric worsened the results for both splits. This limits the specific attribute and method of analysis; it does not refute the role of interactions in tissue. Indicators are not a measurement of consciousness. Without additional verification of the timing of the original records, the calculation should be read as an exploratory association analysis rather than a prospective prediction or causal test. Full group errors and exceptions are stored in chapter31_result.json.

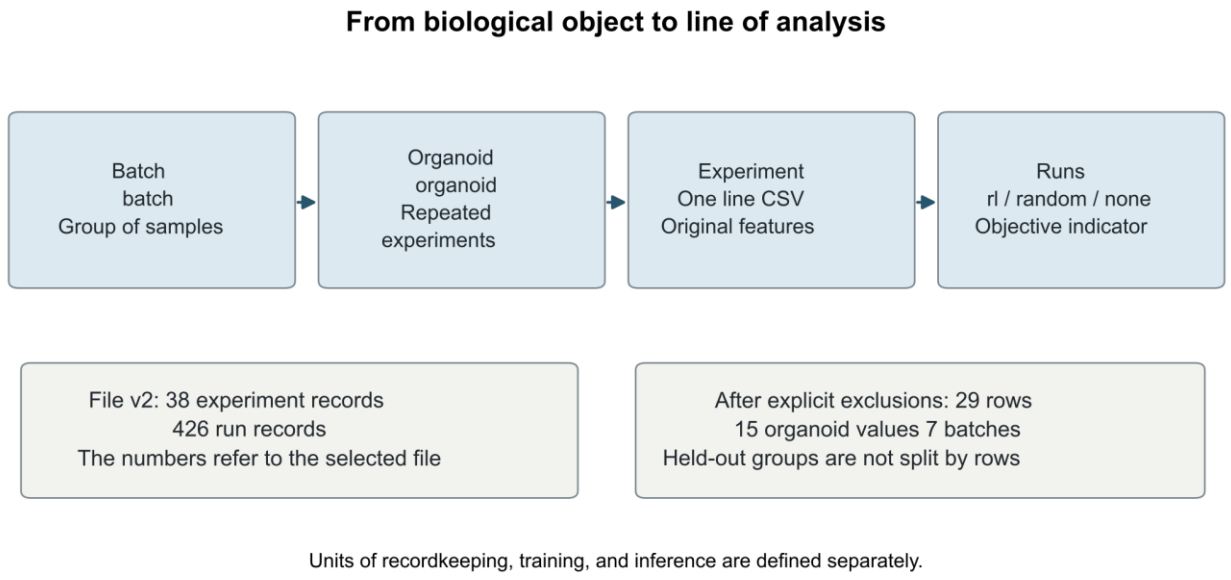

Figure 78. Levels of accounting in biological data. Accounting levels in the verified file aggregated_experiment_data.pkl v2 [47]. Counters describe the contents of a file and do not replace the sample size of a published work. In the study table, the unit of the row is experiment; separation by organoid and batch is checked separately.

Practice 31 With open data

Why do it: Reproduce feature preparation and split-by-sample comparison of two models.

What you will need: Grouping of observations, distinction between correlation and causal intervention, basic and extended forecast. For basic analysis, the attached table is sufficient; Loading the source array requires additional libraries and is described separately.

Time reference: 45–60 minutes. Can be done intermittently; This is not a speed standard.

1. Open bio_provenance.json and bio_features.csv. Match 38 experiment records, 426 runs, and 16 organoid field values with the set description, which lists the other totals.
2. Check the formulas of two characteristics on the parsed line. In the source file, missing entries and infinite values are not replaced by zeros.
3. For a given analysis, 29 eligible records, 15 organoid values, and 7 batches are used. The reasons for excluding the remaining rows are stored in the table.
4. Run python teaching_lab.py --chapter 31. Compare the discharge frequency model with the STTC-augmented model when laying a whole organoid and then a whole batch. Maintain the results even if they get worse.

What to keep: Reproduced analysis describing row, grouping, exceptions, and limitations of statistical inference.

Hint: A deposited organoid from a familiar batch and a completely new batch are different transfer conditions. Before a prospective conclusion, a separate check of the time of obtaining the initial characteristics is necessary.

Exploratory continuation: From the primary timestamps, develop a forecast of the actual subsequent execution change. Determine the sign, outcome, and grouping in advance. For a causal explanation, prepare a separate intervention with a professionally chosen method of intervention.

Analysis and answers - Appendix F, chapter 31, p. 295

Respect the structure of observations

Multiple cycles of a single organoid provide a detailed history of that specimen, but do not evolve into multiple independent organoids. Therefore, randomly dividing all lines into training and testing parts can test the recognition of an already familiar pattern. If the question concerns transfer to new samples, entire samples need to be separated. In the case of common donors or batches, the transfer boundary requires additional clarification.

Another question concerns the dynamics within a particular sample. Here you can train a prediction on an early part of the record and test it on a later one, explicitly limiting the output to this task. The two diagrams answer different questions. You cannot use easy testing within a sample and then declare transfer to a new biological system.

Check also the contribution of the external algorithm. If it changed the feedback signal simultaneously with the tissue change, the resulting dynamics belong to the entire circuit. Useful analysis considers the algorithm's actions as part of the data. Otherwise, the improvement provided by external selection of influence may be erroneously attributed entirely to the neural network.

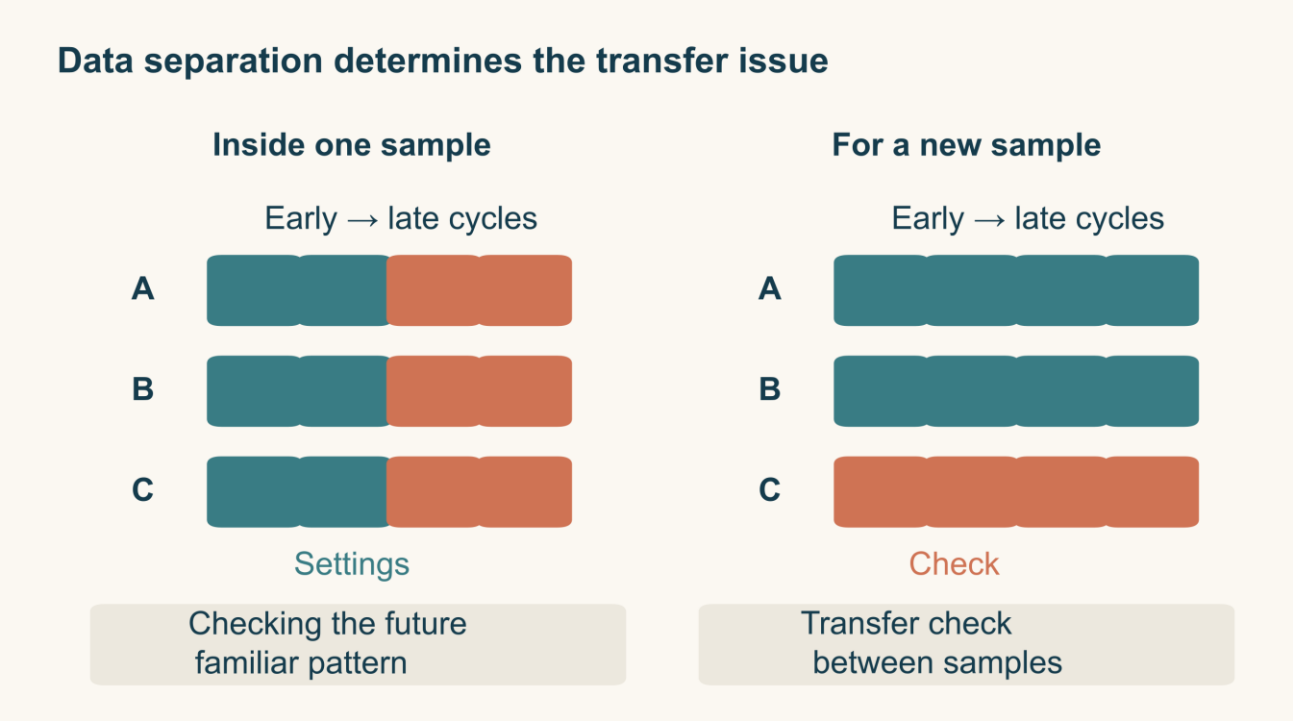

Figure 79. Temporal and group separation of biological data. The training diagram with conditional patterns A–B shows two different questions: prediction of late recordings of a familiar pattern and transfer to a completely delayed pattern. Cycles of the same sample remain linked by observations; with common donors or parties, this level of grouping will also need to be taken into account.

From digital model to intervention design

After comparing the two digital mechanisms, prepare a biological validation project. Specify what difference should occur, what metric would detect it, and what alternatives would produce the same result. The method of actual influence is selected by specialists taking into account the system under study. What is important here is causal logic, not a universal instruction for cultivation or stimulation.

For example, a design can match a stored loopback with suitable input playback control. In this case, it is necessary to take into account the number of impacts, their distribution and the condition of the sample. If control receives a different total input volume, the difference cannot be clearly attributed to closure. The digital stage helps to see such mixtures in advance.

A separate section of the project is devoted to physical organization. If our hypothesis requires mutual causal relationships, we need reasons to believe that they exist in the tissue and circuit under study. A successful statistical measure of connectivity does not automatically establish the desired causal structure. Appropriate interventions or more limited inference about the observed relationship will be required.

Professional working conditions

The ISSCR recommendations serve as one of the guidelines for stem cell research. They are not a universal law that supersedes the applicable policy of institution and country. Update 2025 concerns stem cell embryo models; the remaining sections retain the edition 2021 of the year. Specific requirements must be established for a specific site and project. [48]

The prepared project must define the origin of the material, acceptable uses, responsibilities for measurements and the procedure for reviewing the study. For a hypothesis about possible experience, it is also necessary to describe in advance what new information will change the assessment of the status of the system. This is part of a clear formulation of the study, not a way to settle the phenomenality debate in advance.

The result of practical work

Prepare a two-page research project. On the first page, indicate the question, the available data, the two mechanisms, and the transfer verification scheme. On the second is the proposed intervention, control, restrictions and a separate argument about the connection with experience. If there is insufficient data, describe the missing fields and explain their role.

The work is complete when the transition from a cellular or network change to a specific task outcome can be traced. The final transition to experience must remain an explicitly formulated hypothesis. It is this chain that allows the biological branch to participate in the overall research goal of the book.

Self-test

1. Why can't the improvement of the hybrid system be immediately attributed to the fabric alone?
2. What is the difference between testing on late recordings of the same organoid and testing on new samples?
3. Does temporary improvement in a task prove subjective experience?

Biological evidence has shown the importance of boundaries and grouping. Now let’s put together a unified procedure for checking the implementation: from the existence of connections to their function and a separate argument about experience.

Chapter 32

Verification of the designed system

Digital and biological material require different tools, but have the same need for precise testing. The presence of a component, its functional contribution, internal access, and hypothesized experience are different statements. In this chapter, we will create a profile of the system so that each statement has its basis. You can start with one fully traceable episode; there is no need to test all of the agent's intelligence at once.

Definition of the subject of inspection

You have assembled an agent with memory, an action model, and a self-assessment module. It solves the problem and communicates when in doubt. Now you need to set the status of the result. Program review, functional ability assessment, and experiential exploration are related but require different evidence. If you combine them into one overall score, it becomes unclear what exactly changed after the next improvement.

Start by describing the system as an operating whole. What processes are performed in one solution? What is stored between solutions? What data is available to each component? Is external language service enabled? Does the report derive information directly from the agent's state or only retell the history of actions? The boundaries of the object being tested determine the meaning of all subsequent results.

Map the existing connections

For each component, write down the input, state, output, and destination of the output. Then follow one specific episode. For example, the uncertainty of the action model has increased; the monitor noticed a decrease in expected success; policy requested additional monitoring; a new observation changed the decision. If additional observation concerns only the target, it may not resolve control uncertainty. The map should show such limitations of the mechanism.

The name of the module does not confirm its role. A component can exist in the code, but never turn on in the selected sequence. Its condition may change, but not affect the decision. Or the influence is completely duplicated in another way. It is therefore useful to distinguish between the availability of a computation, its actual activation, and its causal contribution to the outcome.

Check the measuring system separately. The true state of the environment and intervention flags must be available to the evaluator, but not bypassing the agent. If the self-assessment module reads a ready-made correct answer, the high accuracy of its messages indicates the quality of the transmission of this information. It does not install independent monitoring of its own solution.

Test the feature under new conditions

Memory requires tasks where the story actually changes the correct action. For the control model, conditions where the connection between command and effect changes. To assess uncertainty—situations where certainty may be wrong. One convenient demonstration does not test all of these abilities at once.

It is useful to separate familiarity from transference. In a familiar problem, the agent could learn a simple rule sufficient for success. The new combination of features will indicate whether it uses the intended structure. But a new condition should not imperceptibly change everything at once. If the noise, the goal, and the available actions change simultaneously, it is difficult to attribute failure to a specific mechanism.

Select a primary result for each feature. The number of requests in itself is not a quality of self-control: you can ask endlessly. The probabilistic prediction is evaluated against the subsequent outcome, and the verification policy is evaluated against benefits and costs. The content of a single confident sentence does not replace the distribution of errors over many episodes.

Prepare a causal test

Let's imagine two educational systems with equally convincing reports. The first forms a message from the source’s own assessment. The second chooses a phrase based on recent public actions. As long as the events are common, the difference may not be noticeable. Let's change the internal assessment of the source while maintaining the external story. If the first system actually uses this estimate, its message should change predictably.

However, interference may create an external sign: an unusual delay, a special token, a different response format. Then the second system also recognizes the condition. Therefore, control must preserve accessible external signs. A dummy intervention and a suitable nonspecific change are also needed to distinguish content recognition from a response to a general failure.

In a research report also posted as a preprint, Jack Lindsay examined the relationship between language model messages and interventions in their internal representations. The acquired abilities turned out to be limited and dependent on

conditions. What is important for us is the transition to causal verification of internal access; it does not automatically infer subjective experience. [49]

Distinguish reading state from using it

An external classifier can be trained to determine the content of a latent vector. This will show that it contains information available to the selected analysis. But the agent may not use it either for reporting or for action. Therefore, external decoding and internal access should have different lines in the research description.

The next question is whether access affects control. Disable the monitor's association with the policy, leaving the monitor itself intact. If the estimates remain correct, but the decisions no longer take uncertainty into account, the contribution of this connection is highlighted. If the behavior does not change, the mechanism may be inactive, duplicative, or a task where it is not needed. You cannot choose a convenient explanation without checking.

Finally, strategy revision requires its own outcome. The agent must change a certain assumption and use the change in new decisions. Explanatory text after an error can be a useful interface, but its presence does not confirm such a causal chain. Comparison with a version that speaks the same text without revision helps highlight the mechanism at work.

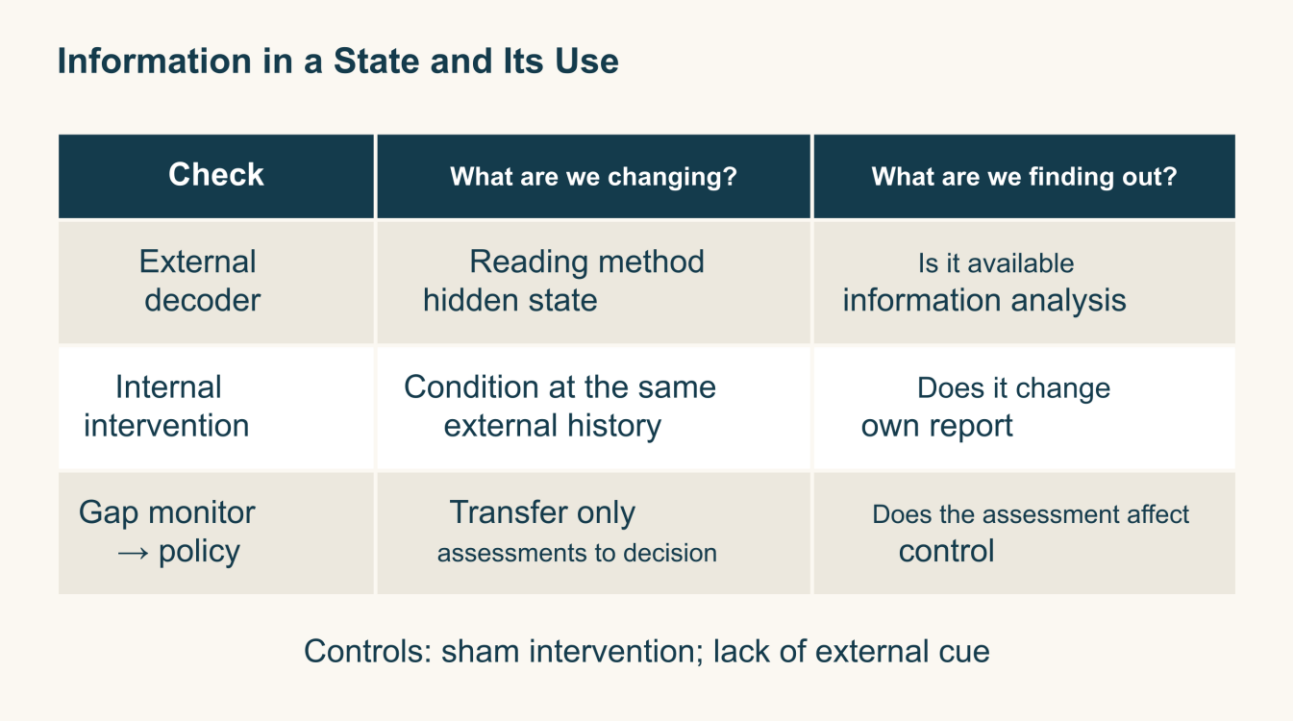

Figure 80. Three different questions about the internal state. External decoding, causal change in self-report, and monitor policy influence require different tests. The internal intervention must preserve external history and have fictitious control, otherwise the system may recognize a secondary indication of the influence.

Build a candidate profile

The theoretical profile begins with a selected position, rather than with a general list of spectacular properties. If the role of public access is being explored, the appropriate organization of distribution and use of content needs to be examined. If physical causal structure is significant, implementation data is required. If the chapter principle 26 is selected, the mutual determination of the distinctions and their role in alternative transformations are checked.

The indicator method of Butlin and colleagues suggests linking system assessment to the requirements of theories. It does not provide a universal scale on which the number of items completed turns into percentage of consciousness. In our work, the theoretical profile preserves the basis and uncertainty of each individual conclusion. [42]

For each requirement, use three descriptive states: implemented and verified; implemented, but the role has not been verified; not implemented or unknown. Then write down which observation can change the status. Such a record is more useful than the total number, because it shows where work with the code is needed, where a new experiment is needed, and where clarification of the theory itself is needed.

From a Profile to a Claim about Experience

A successful functional profile does not complete the argument. It is necessary to justify why the selected properties relate to the presence or quality of the experience. The results of the chapters on natural experience and the principle of correspondence formulated in the research section are used here. It is important not to consider the fact that one’s own requirements have been met as independent confirmation of their sufficiency.

Let us assume that the system satisfies our working organization of mutual distinctions. Then we can assert that a certain realization of the supposed sufficient condition has been created. The next question concerns the condition itself: does it explain independent data, withstand counterexamples and carryover between implementations? These reasons must be presented separately from the technical report.

If a theory allows for experience without explicit self-report, the absence of persuasive speech does not automatically disprove it. If a theory requires a certain access, good speech without the necessary causal organization does not support it. The test should follow the content of the theory, and not the impression of communicating with the system.

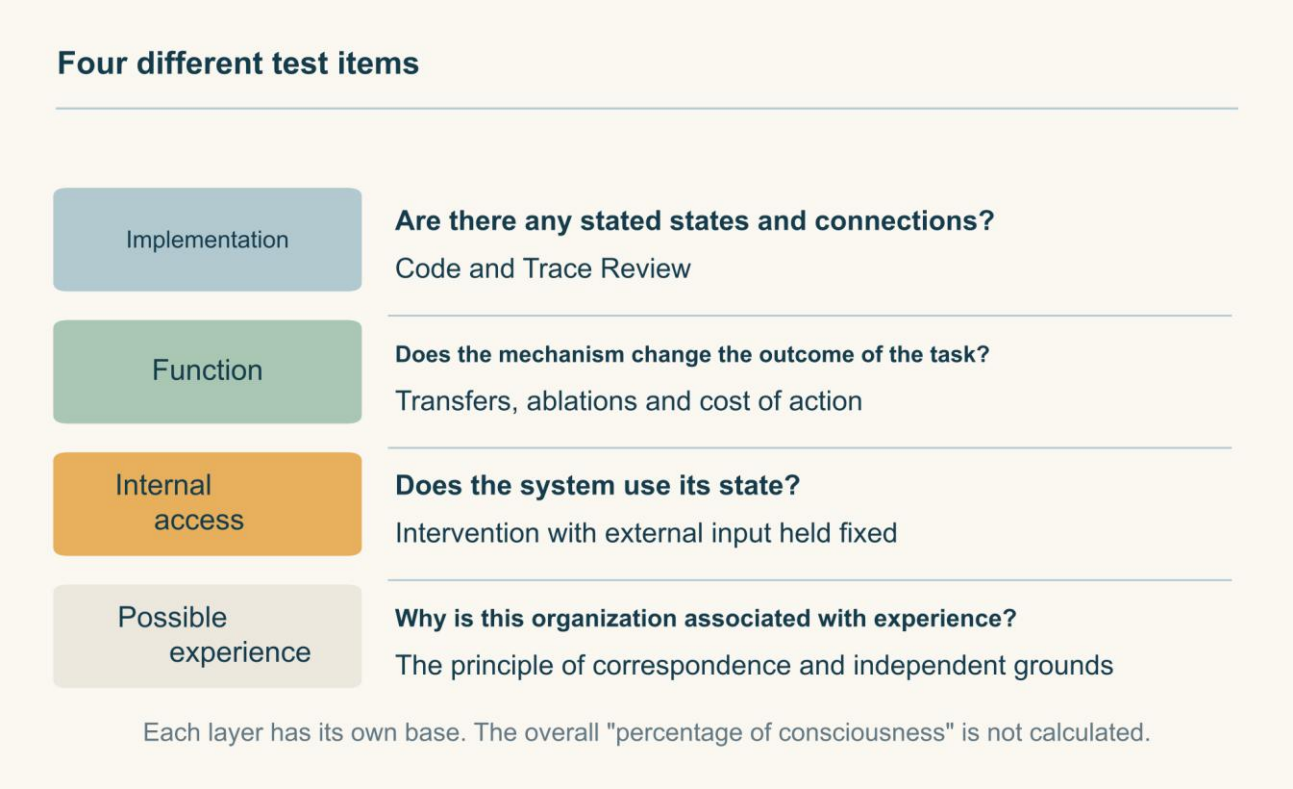

Figure 81. Candidate Screening Profile. The methodological table distinguishes between the presence of a software implementation, its functional contribution, the internal use of information and the argument about possible experience. Evidence from one layer does not replace evidence from another; the external state decoder does not itself establish internal access, and the sum of the requirements fulfilled is not a percentage of consciousness.

Interpreting a Result That Does Not Distinguish Models

Imagine that the full version and the version without the strategy revision gave the same result. First of all, check whether the conditions for triggering the review have occurred. If at no time, the experiment did not exploit the differences between the versions. The statement “reflection is not needed” is too strong, as is the statement “reflection works because the module is written.”

Then find the problem where the rules actually diverge and print the expected difference before running. If the results are the same again, investigate possible duplication and measurement accuracy. The need for a component is established relative to a specific implementation and task. Another system may provide the same result in a different way.

Such analysis has direct value for designing experience candidates. It prevents the accumulation of decorative components and helps to select mechanisms with a proven functional contribution. The phenomenal hypothesis receives a more precise subject, even if the final answer about the presence of experience has not yet been found.

Four questions for one status record

In the M2 model, record A actually exists: this is verified by the program state. Its influence on repeated A is checked by comparison with the remote history at the previous input. Recording access on a separate monitor will require checking the read channel; The presence of a record in itself does not create it. Finally, if the model gives the required numbers, it is still necessary to explain why this organization relates to the experienced content. Thus, one log can support several different checks, but each transition requires its own monitoring.

What has now been established

A profile helps avoid summarizing disparate attributes. The unknown property status indicates a specific next job instead of a general impression of awareness.

Practice 32 Perform an audit of one episode

Why do it: Parse one episode of implementation into four independent statements.

What you will need: Causal map, ablation as disabling a specific component, difference between decoding and using state. Keep the same available information across the versions you compare.

Time reference: 30–40 minutes. Can be done intermittently; This is not a speed standard.

1. Select repeated A in the source model or one agent step. Write down the input, old state, change, output, and output recipient.
2. State the connection to be disabled and the expected result before testing. Make sure it's the one that turns off.
3. Create four profile lines: implementation, function, internal access, possible experience. For each, indicate whether it was verified or unknown.
4. Offer one task where the feature is truly needed, and control that preserves comparable information and resources.

What to save: Single episode map, causal test and profile with separate statuses.

Tip: An external classifier may extract content that the agent does not use. For internal access, a recipient of information is needed inside the circuit under study.

Research continuation: Check the text report when the internal state and external history change independently. Compare with an explanation formed only from public actions.

Analysis and answers - Appendix F, chapter 32, p. 295

Self-test

1. Does the presence of the Monitor class in a program prove working self-monitoring?
2. How is external state decoding different from internal access?
3. Why doesn't meeting all self-selected requirements prove that those requirements are sufficient for the experience?

We have learned to check a specific implementation. Now let's clarify the rules of formalization and find out whether the selected analysis is even capable of distinguishing between the mechanisms that we want to compare.

Part 9

Research work

Chapter 33

Formalization and simulation

Testing a program does not end with the correct calculation of one example. Two sets of parameters may produce similar results, and different models may be indistinguishable in the chosen environment. In this chapter we formally separate the generator, available observations, forecast and measurement. Then we will perform the restoration of parameters and models on artificial data, where the generating mechanism is known. This is the preparation of an analysis that will later work with an unknown cause.

Formalization makes commitments visible

The verbal model allows for useful ambiguity early on. But to test it you will need to decide what exactly happens between the two events. Is history preserved while the source is missing? Does the probability of the underlying cause change every moment or only with a new observation? Does the contradiction update one record or the entire system? These decisions can change the prediction, although the brief description of the model remains the same.

Formalization does not necessarily mean making mathematics more complex, but rather eliminating such ambiguities. The input, state, transition, and measurement must form a procedure that another researcher can reproduce. The program helps you check the consequences of the rules. However, the agreement between the code and the specification and the suitability of the specification for explaining the experience remain distinct issues.

Define the learning environment separately from the model

Let's expand the example of two sources A and B. Each has a hidden binary state: a common cause or different causes. In a learning environment, the state of the source is preserved with probability 0.9 the next time it is accessed. The observed positive sign appears with probability 0.75 for a common cause and 0.25 for different causes. This is a constructed environment, not a description of a physiological mechanism.

The experimenter sets hidden states and observations. The model receives only the source and the feature. Therefore, the prediction of the next symptom and the assessment of the underlying cause can be measured separately. If the model sees the hidden cause directly, the filtering task will disappear. The separation of the environment and the observer must be maintained even when they are written in the same file.

The retention rate of 0.9 corresponds to the retention rate of 0.8 from chapter 25. After a possible change in reason, the previous belief shifts to 0.5 according to the rule that we have already used. Now this rule has a certain meaning in the data generator. A different environment might require different dynamics.

Determining the Predicted Outcome

The belief 0.8 means the probability of a common cause. It is not equal to the probability of the next positive sign. If 0.75 is reliable, the next sign may turn out to be negative even with a common cause. The prediction of a positive sign is obtained by mixing two possibilities: the probability of a common cause is multiplied by 0.75, the probability of different causes is multiplied by 0.25; the results add up.

```
def predict_positive(prior_common, reliability=0.75):
    return (
        prior_common * reliability
        + (1.0 - prior_common) * (1.0 - reliability)
    )
```

Given the belief 0.8, the prediction of a positive sign is 0.65. This is a verifiable numerical difference. If one compares a belief about a cause with a trait's frequency directly, the model may erroneously appear to be poorly calibrated. Therefore, the name of each output number must indicate the event whose probability it represents.

The human assessment of belonging will add another layer. It is neither a true causal cue nor a subsequent sensory cue. The connection between the internal state and such an assessment requires its own measurement model. Formalization helps to notice these transitions before analyzing the results.

Record the model time

In our simple M2, retention is applied the next time a particular source is accessed. If A remains absent for a long time, its record does not weaken at each intervening event from B. This is a convenient rule for a teaching calculation. A hypothesis about human memory, however, may require dependence on elapsed time.

The second version will attenuate the recording for each time step. With the coefficient 0.8 and two skipped steps, the influence of the previous deviation will be successively multiplied by 0.8 twice. Now the duration of absence will become an independent predictor. The two versions can be compared by presenting the same number of events at different intervals.

Don't quietly correct the first version after seeing an inconvenient result. Changing the model time means a new hypothesis. Keep the old rules, describe the reason for the revision and get new predictions. This is especially important for connection with experience: subjective continuity should not be explained by an arbitrarily changing concept of time.

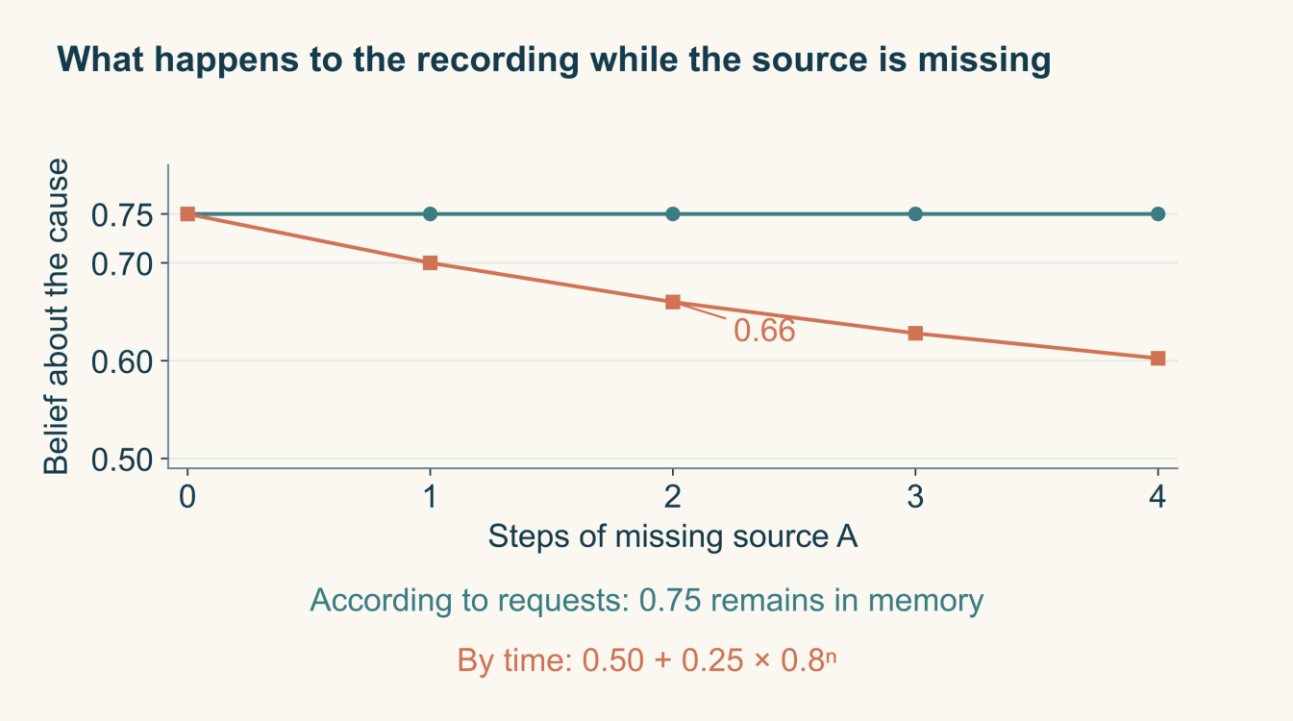

Figure 82. Two clearly different models of memory timing. Accurate training calculation when initially recording 0.75 and holding 0.8. The saved state is shown during the absence of A, before a new access and before a new observation: in two skipped steps, the temporary version comes to 0.66, and the access version stores 0.75.

Check parameter values

Two mechanisms can produce the same final number. Belief 0.75, if the positive attribute 0.75 is reliable, is updated to 0.9. The initial belief 0.5 with the reliability of 0.9 is also updated to 0.9. Whether a high value is due to a strong story or strong new evidence cannot be determined from a single summary response.

To distinguish the parameters, other conditions are needed. You can independently evaluate the discrimination of the current feature, change the length of the history, and add contradictory events. Then check whether the different sets of parameters

remain practically indistinguishable. Parameter uncertainty is a result of the analysis, not an obstacle to be hidden by choosing one nice set.

Try recovery on artificial data. Set the parameters, generate the sequence, then have the analysis find them without access to the original values. If a procedure routinely confuses retention with reliability, human subjects research will face the same predicament. Increasing the number of similar events may not eliminate structural ambiguity.

Check model recovery

Now generate separate datasets with M0, M1, and M2, adding the predefined variability in observation. The analysis must compare every model for each dataset. Ideally, it will favor the mechanism that generated the data, but this cannot be assumed. The conditions may barely distinguish the competing models, or the selection rule may systematically favor the more flexible model.

Reconstruction of parameters and models is discussed in the methodological work of Wilson and Collins. Our specific generators and sequences serve as a stand-alone tutorial on this logic. A successful test here demonstrates the ability of the assay to perform under known conditions; it does not prove that real data is generated by one of the models considered. [43]

It is useful to add a fourth generator that does not match any candidate. Then you can see how the procedure selects the best of the incorrect models. Even a convincing victory in a limited set of competitors does not mean that a valid mechanism has been found. Testing the absolute quality of the forecast should complement the relative comparison.

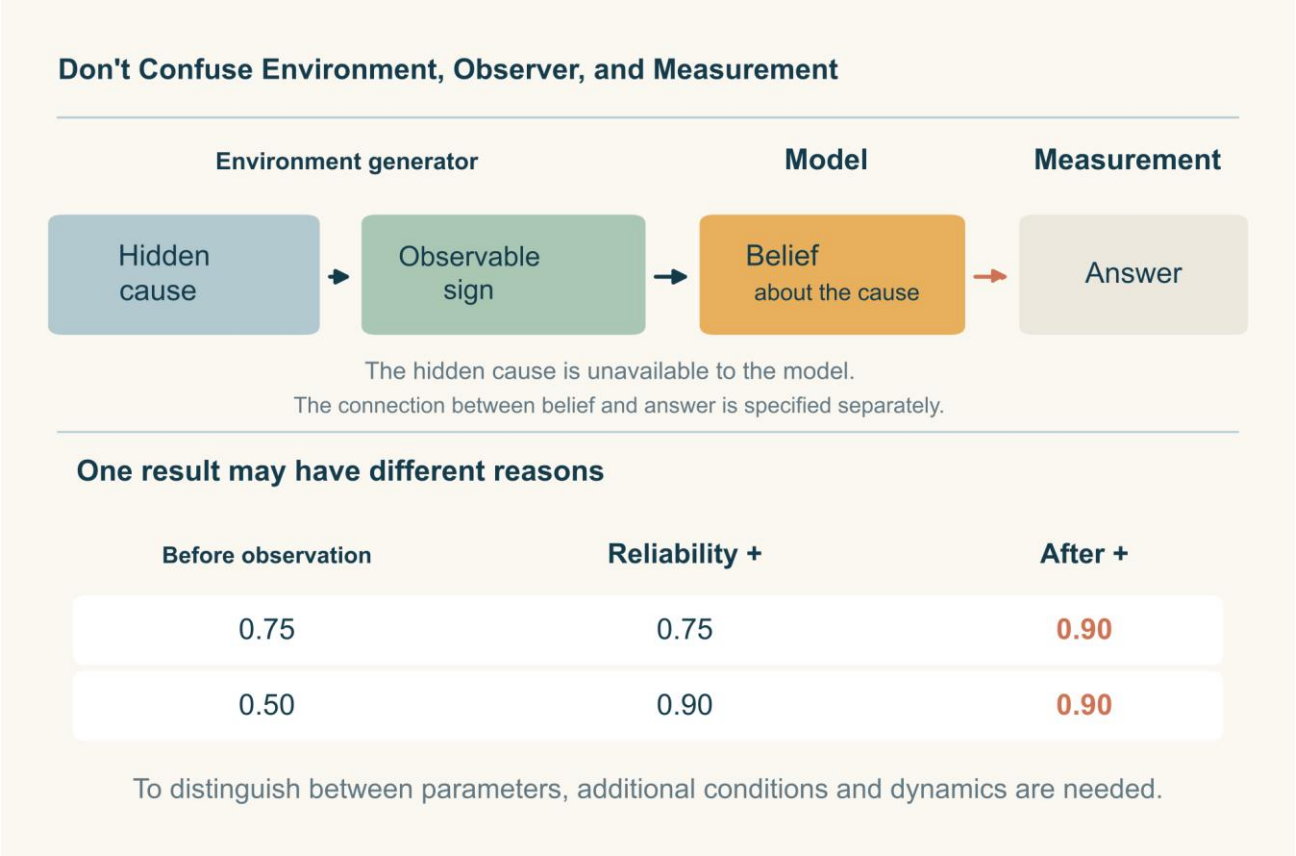

Figure 83. The limit of available information and the ambiguity of parameters. Training diagram and accurate calculation. The upper part separates the latent cause in the environment generator from the available observation model; The dotted line denotes a separately specified response generation model. In the lower example, positive evidence with reliability 0.75 under preliminary persuasion 0.75 and with reliability 0.90 under persuasion 0.50 gives one result 0.90. Here, the pre-belief is given immediately before observation: no additional memory decay is applied. One total does not determine both parameters.

Maintain reproducibility

For each run, record the rule version, parameters, initial state, and order of events. When generating randomly, fix the initial values of the generators. But one seed is not enough if the sequence of calls to randomness has changed. Additional observation should not inadvertently shift the entire future scenario of the compared version.

Separate condition generation and agent decisions. Potential observations at each step can be prepared in advance, and the agent's action will determine which one is discovered. If the action physically changes the future environment, this change must be part of the causal model. Comparability does not mean an artificial prohibition on the consequences of an action.

Also save the raw results and the summary program. Otherwise, fixing one filter may imperceptibly change the numbers in the text. In the final set, the reader should see how the sequence produced predictions, the predictions - errors, and

the errors - the conclusion. This chain is more important than the decorative complexity of the interface.

From simulation to exploratory inference

Simulation can establish that a mechanism is realizable, what consequences it has, and under what conditions models differ. It can also show the impossibility of the desired result given the chosen rules. But the artificial data generated by the model itself does not independently confirm its applicability to a person or biological network.

For the hypothesis of chapter 26, additional description of the execution is required. If computational organization is significant, show that the program actually implements it, including responses to interventions. If certain physical interactions are significant, justification at this level will be required. The simulation graph does not replace the question of what process is actually occurring in the medium.

Executable recovery procedure

In the extended workshop, the generator M0 creates independent binary features with probability 0.5. The M1 generator stores one hidden cause for the entire sequence, the M2 generator stores a separate cause for each alternating source. When contacted, the cause remains with probability 0.9, and a positive sign appears with probability 0.75 for reason 1 and 0.25 for reason 0. The initial causes are equally probable. Thus, the generating environment is specified independently of the analyzing model.

Model recovery uses 30 independent sequences of 120 events for each of the three generators. Analyzers M0, M1 and M2 receive the same source labels and features, but not hidden causes. In each trial, the probability of a positive attribute is first stored, then its meaning is discovered and the belief is updated. The main criterion here is the mean square error. If the results are exactly equal, a draw remains; randomly assigning a winner does not add distinguishability.

To restore the parameters M2, a predetermined retention grid from 0 to 1 with a step 0.1 and a reliability grid from 0.55 to 0.95 with a step 0.05 is used. For each pair, the error of the complete trajectory is calculated on known artificial data. This small search makes neighboring nearly equal solutions visible. It is a training procedure and not a universally best method for estimating parameters.

Control with one source must give equality to M1 and M2. Additionally, all series where the selected analysis did not favor the generating mechanism are saved. The matrix of such choices shows the ability of this particular amount of data and

criterion to distinguish between these particular generators. For a different environment, different noise or different mesh, the test is performed again.

After checking the instrument, you can set the main experiment of the chapter to 34. The use of known artificial causes here is legal for the evaluator who verifies the restoration. Transferring these reasons to the predictive candidate itself would defeat the purpose and make success uninformative.

One outcome does not determine the origin of confidence

Belief 0.75, when the positive attribute 0.75 is plausible, is updated to 0.9. The original 0.5 with the reliability of 0.9 also gives 0.9. A finite number alone is not enough to restore the power of past and new evidence. Therefore, analysis needs a trajectory, contradictory signs and different conditions. Testing on data from a well-known generator shows where the procedure confuses parameters and where it confuses entire models. If the competitors are computationally identical, such a test should detect a tie rather than artificially assign a winner.

What has now been established

Simulation allows one to explore the consequences and discriminability of rules. It is not independent evidence of their applicability to human experience. Now let's check the comparison tool itself.

Practice 33 Explore mechanism uncertainty

Why do it: Reproduce recovery on known generators and find the area of ambiguity.

What you will need: Update M0-M2, difference between the probability of a hidden cause and the probability of a sign, a unit of independent repetition. The workshop added an executable example of comparison and recovery.

Time reference: 40–60 minutes. Can be done intermittently; This is not a speed standard.

1. Run python teaching_lab.py --chapter 33. Generators and parameter grid are described before practice; all predictions are recorded before the current feature is opened.
2. Consider the errors M0, M1 and M2 on each series, then the selection table for a known generator. Save all model classification errors.
3. For M2, compare several retention and reliability values. Find neighboring combinations that are close by mistake and check intermediate states.
4. In the single-channel condition, check the equality of M1 and M2. Compare the time by calls and calendar steps in the sequence with missing A.

What to keep: Restoration result, known equality control, and one condition that enhances parameter discrimination.

Hint: Do not discard series in which the candidate that did not generate the data won. They characterize the chosen procedure and the amount of information.

Exploratory follow-up: Add a generator that does not match any candidate and evaluate the absolute quality of the best model. Test the recovery under a different variability and number of observations by committing to a new plan before running.

Analysis and answers - Appendix F, chapter 33, p. 296

Self-test

1. Why is the probability of a common cause different from the probability of a positive attribute?
2. Does the restoration of M2 on the data of M2 itself prove its truth for humans?
3. Why does changing the forgetting rule require a separate version of the model?

The limitations of the analysis itself are now known. In the next chapter we will document how it will be applied to new data so that the result does not select its own validation rules.

Chapter 34

Protocol before data collection

The preliminary plan links the question, future data, and outcome decision. It is needed not only to protect against arbitrary selection of a successful schedule: it makes it clear whether the procedure is checking exactly the desired mechanism. Here we will prepare an executable protocol for comparing common and separate history. His logic then carries over to human and biological projects with their additional levels of grouping and measurement requirements.

Why Decisions Should Be Described in Advance

Imagine that you have received several graphs. On one M2 is better than M1, on the other the difference has almost disappeared. You can select the first chart, change the time window and explain why this particular indicator always seemed to be the main one. Such a story may sound convincing even to the author himself. The preliminary protocol records what you actually expected and intended to check.

Pre-registration does not preclude investigation of unexpected results. It separates testing a pre-formulated expectation and creating a new hypothesis based on the data seen. This distinction lies at the center of Nosek et al.'s methodological proposal. Transparency of decisions is essential to our work, not the ceremony of document placement. [10]

Below is a training sample protocol. It does not constitute a report of registered or completed research. You can fill it out for your own future review, saving the date and version. If the data has already been studied, this should be indicated and new independent material selected for the confirmatory step.

Start with one main question

Our question is whether separate source histories predict their next features better than a shared history. This is the computational stage of the project introduced in Chapter 33. It does not directly test the human sense of ownership or establish that the program has experience. It does, however, define a mechanism precisely enough for later comparison with more complex data.

The main hypothesis states the advantage of M2 over M1 in an environment with two independent stable sources. The control hypothesis states the agreement of these models when the source is the same and the retention rules are the same. M0 serves as a guide without transferring history. M1I, implemented by the same

separate entries, is the same as M2 and is not considered a separate empirically distinguishable alternative.

Also write down the reasons for the hypothesis: M1 mixes evidence from different sources, and M2 keeps them separate. This is the mechanism of hypothesized advantage. The words “smarter model” add nothing to it. If conditions do not create conflicting histories, the expected difference may be very small.

Completely define future data

The training protocol plans two hundred independent sequences of one hundred and twenty events. These numbers indicate the reproducible volume of the exercise and are not a standard for the sufficiency of scientific sampling. Half of the sequences contain one source, half - two. In the two-source version, the appeals alternate. The initial hidden states are chosen with equal probability.

Each source, the next time it is accessed, retains a hidden state with probability 0.9. The probability of a positive observation is 0.75 for a common cause and 0.25 for different causes. The models know these parameters, but not the hidden states. This option tests the mechanism under favorable certain conditions; Robustness under incorrect assumptions will be a separate study.

Before generation, secure the randomness algorithm and the list of initial values. It is convenient to use sequential values from one to two hundred and match the condition to the number in advance. Do not select a successful seed after viewing the results. If a mixed distribution of conditions is required, the rule for this distribution is also fixed before launch.

Determine the forecast and when to record it

Before each event, the model calculates the probability of a positive sign. First, the old belief is updated taking into account a possible change in the cause, then an observation forecast is obtained from it. Only after saving the forecast is the current sign revealed and the belief changed. This order prevents the answer from being used in predicting the answer itself.

The main indicator is the mean square error of the probabilistic forecast, usually called the Brier score. For a positive event the actual value is one, for a negative event it is zero. The actual value is subtracted from the forecast and the difference is squared; the errors are then averaged. A smaller value means a better probabilistic forecast in the chosen sense.

For example, the forecast 0.7 before a positive event gives the error 0.09, and the forecast 0.4 gives the error 0.36. These numbers illustrate calculations and are not the results of future testing. The indicator refers to an observable characteristic. It

cannot be called a measure of qualia or directly used as a scale of the validity of a person's sense of self.

Capture the Main Comparison

For each sequence, calculate the difference between the errors M1 and M2. A positive difference means the advantage of M2. The main outcome is the average pairwise difference in the two-source condition with uncertainty estimates from independent sequences. Events within one sequence are connected by history, so they cannot be considered the same number of independent repetitions of the entire study.

In the single source condition, an exact match of the selected realizations M1 and M2 is expected. This is control of program logic. If the results are different, you should first look for differences in inputs, updates, or initial states. Such an incident should not immediately turn into a surprising discovery about memory.

Comparison of M2 with M0 and graphics for individual steps can be indicated in advance as additional. If a new time window appears after reviewing the data, its analysis remains useful but takes on exploratory status. The main comparison does not change retroactively just because the additional one turned out to be more beautiful.

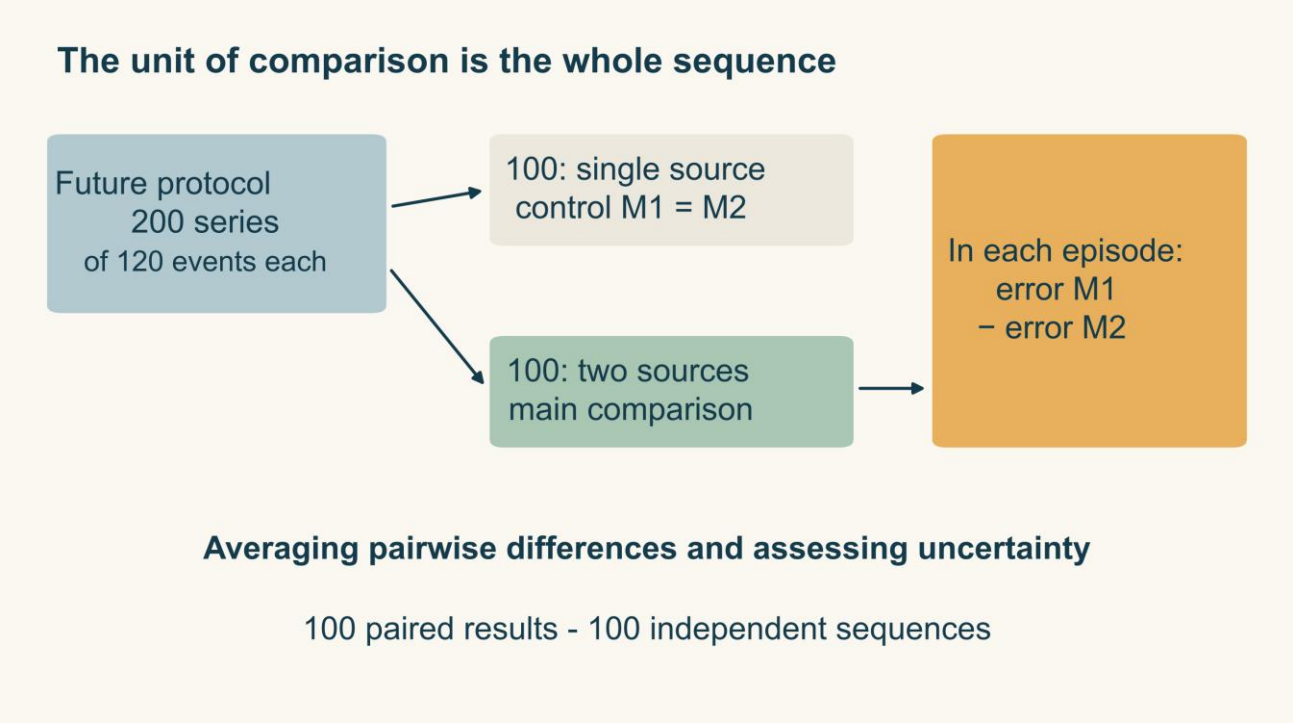

Figure 84. Data structure of the preliminary protocol. The numbers are taken from the future training protocol of chapter 34 and indicate the volume of the exercise, and not the norm of scientific sampling. For the main comparison, the uncertainty is estimated from one hundred independent sequences with two sources, maintaining the pairing of the results M1 and M2.

Specify the criterion for a meaningful result

A positive mean difference can be so small that it has little effect on the use of the model. Therefore, in addition to the direction, specify the effect size of interest or the required estimation accuracy. In a teaching project, you can pre-select the difference Brier 0.01 as a reference point for discussion. This is a conditional threshold of the exercise, and not an established universal significance limit.

For a serious project, the threshold is derived from the problem: what gain justifies the complication, what difference can competing explanations make? The simulation then helps select the amount of data for the desired accuracy. If the predetermined budget is insufficient, the conclusion should preserve this uncertainty, and not replace it with “no effect.”

You cannot set a stopping rule to “add data until the model wins.” You can choose a fixed volume or a reasonable sequential procedure in advance. In both cases, the rule is included in the protocol. The reasons for the forced shutdown and the actual volume should remain visible in the final report.

Prepare error handling rules

In a digital problem, it is necessary to determine in advance which results are technically impossible: a probability outside the acceptable range, the absence of an event, a discrepancy between the scenarios of the compared models. Such a case requires correcting the implementation and re-running it while maintaining the version. Eliminating an inconvenient but acceptable result does not correct the error.

In a study of a person or a biological network, there will be other reasons: an incomplete procedure, loss of recording, violation of a predetermined quality of measurement. The rules need to be related to the suitability of the data, not the direction of the effect. For biological data, the unit of independent comparison and available sample information are additionally recorded.

Indicate who makes the decision to exclude and when. If a decision depends on knowledge of the group or outcome, this should be disclosed. The protocol does not make the study error-free; it allows you to evaluate how decisions may have affected the conclusion.

Leave Room for Discovery

Let's say M2 especially benefits after returning a source with a contradictory history. This figure was not a major prediction, but may suggest a new mechanism. Write it down, check the technical explanations, and formulate the next

discriminating question. Then prepare a new sequence where the updated idea runs the risk of being wrong.

With this approach, research analysis does not become second-rate. It does another job: it creates the content of the future check. The problem arises only when the two stages are mixed, when the discovered difference is presented as expected in advance and thereby inflates the strength of the evidence.

How one question defines analysis

The question of the usefulness of individual stories requires at least two distinguishable sources. For each independent scenario, you can calculate the average error of shared memory and separate memory, and then their difference. A positive difference means an advantage for M2. If you average all events as independent repetitions, the overall history within the scenario disappears and the accuracy may be overestimated. Therefore, in the training protocol, the repeat unit is the whole sequence, and the single-channel environment is the equality control.

What has now been established

The plan should allow the main comparison to be made without selecting a metric after viewing the graph. At the same time, it maintains a separate place for subsequent discoveries and new hypotheses.

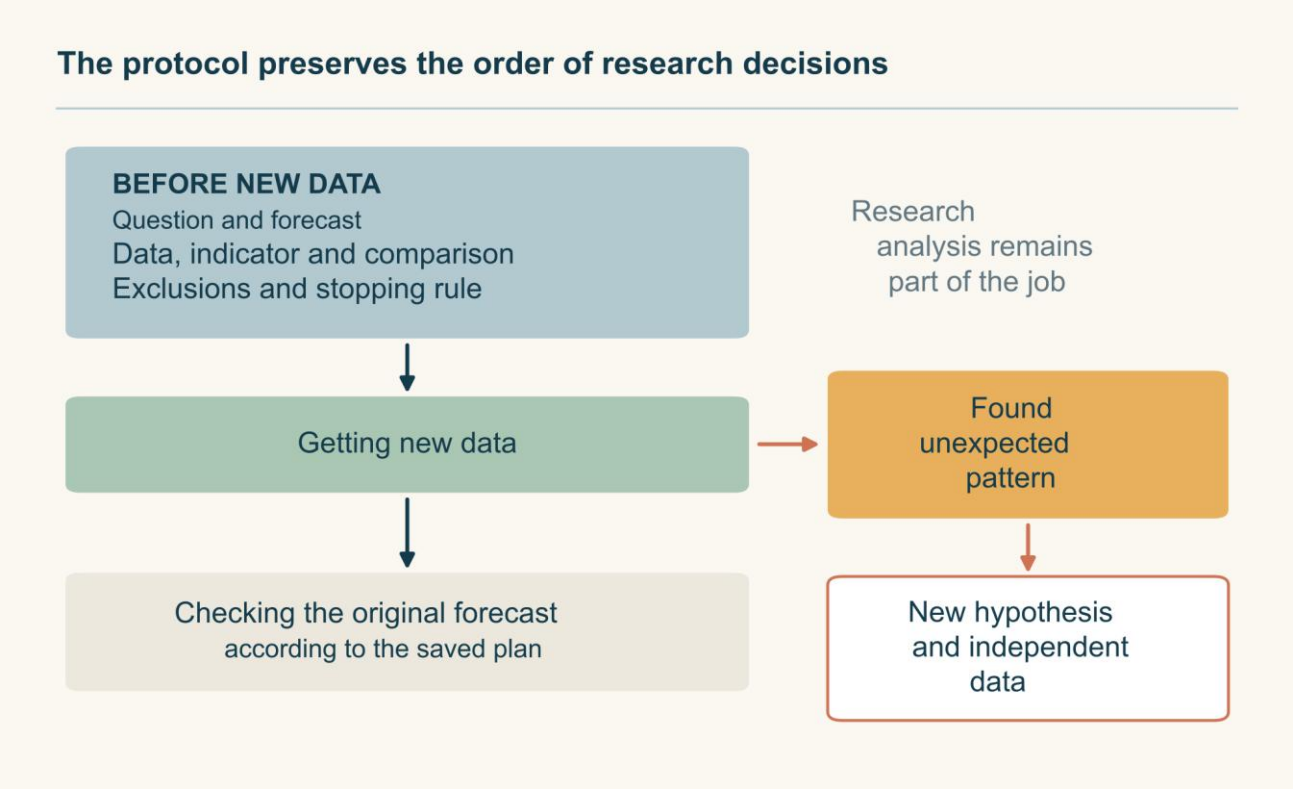

Figure 85. Preliminary planning and subsequent opening. Methodology: A pre-stored protocol defines the confirmatory test. An unexpected pattern in the data obtained can give rise to a new hypothesis and the next project with independent data. This branch is useful, but it does not turn the effect found after analysis into a previously predicted one. The diagram does not report the registration or experiment performed.

Practice 34 Complete your own protocol

Why do it: Draw up a protocol that another person can execute without making additional decisions for the author.

What you will need: Generator and analysis of chapter 33; main indicator, paired comparison, and independent replication unit. Distinguish between new supporting material and already reviewed results.

Time reference: 35–50 minutes. Can be done intermittently; This is not a speed standard.

1. Use the protocol_source_history.txt sample in the workshop. Ask the Brier pairwise difference question for M1 and M2 in a two-source environment.
2. Record 200 independent series of 120 events: 100 with one source and 100 with two. Name the generation rules, parameters, seed and forecast moment.
3. Save the main indicator, the conditional content landmark 0.01, the unit of analysis, the processing of technical errors and the stopping rule until a new start.
4. Indicate what data you have already seen and list the search analyzes separately. When doing your own new research, use new scenarios.

What to keep: A dated version of the full protocol and a clear boundary between the confirmatory and exploratory stages.

Hint: The protocol recorded after familiarization with the result does not become preliminary due to a change of name. Save the origin of solutions.

Research follow-up: Prepare an appropriate protocol for actual measurements, adding eligibility criteria and participant, sample, donor, or lot levels. Justify the required precision in relation to the question.

Analysis and answers - Appendix F, chapter 34, p. 296

Self-test

1. Does pre-registration prohibit unexpected testing?
2. Why does the forecast need to be saved before opening the current feature?
3. Is two hundred sequences a universally sufficient amount of research?

After saving the plan, the results will appear. Chapter 35 will show how they change the output when the advantage is small, the component was not included, or the best candidates still make mistakes.

Chapter 35

Review and revision

The study should end with a decision grounded in its result. Sometimes the data will support a mechanism; sometimes they will constrain the effect size; sometimes they will fail to distinguish the competing explanations. We begin by reconstructing the order of available information and the known controls, then estimate the effect and its uncertainty. Only then do we revise the explanation. This sequence helps us avoid mistaking a technical improvement in a number for scientific progress.

The result should change our expectations

The study ends not with a number, but with a clarification of what is now permissible to assert. A positive result may support a mechanism, reveal a useful engineering opportunity, or show only the advantage of one implementation. A negative result may invalidate a specific prediction, reveal task insensitivity, or remain undefined. Each case requires its own next step.

Let's take the project of separate memory of sources. We expected that M2 would be better than the shared memory M1 in predicting features of two stable sources. After the launch, you should ask what exactly was compared: the same observations, the same forecast time, the same update capabilities? Only then can the size of the advantage and its relation to the original idea be discussed.

First restore the causal sequence

The most convincing mistake is the one that improves the outcome. Imagine that the program first updated a belief based on the current attribute, and then saved it as a prediction for the same attribute. The forecast error will be reduced, but the model will use the answer it was supposed to predict. A large win here indicates a violation of the procedure.

Select several events and restore them completely: old state, available information, preliminary forecast, discovered sign, new state. Check not only the final graph, but also the order of these transitions. For an active agent, add the action, its cost, and the additional observation actually received. It is this record that allows you to establish what information the system had.

Then perform the test for which the answer is known from the definition. In a single-channel environment, our M1 and M2 must match. With zero retention, all versions forget the previous state. Violation of these equalities requires an explanation in the implementation rules. Correcting the error is accompanied by a

new version and re-obtaining the results; the previous numbers are saved as canceled.

Estimate magnitude and uncertainty

Let the difference between errors M1 and M2 in four artificial sequences be 0.018, 0.006, −0.004 and 0.012. These are specially invented educational values, and not the result of a completed experiment. The average difference is 0.008: on average M2 is better, but one sequence gave the opposite result. One sign of the average is not enough for a meaningful conclusion.

In the protocol of the previous chapter, the difference 0.01 served as a reference point. Our illustrative average is less than this benchmark. Moreover, four sequences do not allow one to declare the absence of an advantage of interest without assessing uncertainty. It is necessary to find out what quantities are compatible with the data and how accurate the estimate is relative to a pre-selected question.

For sequences with a common internal history, the unit of repetition remains the entire sequence. You cannot turn four runs of one hundred events into four hundred independent repetitions of the mechanism. If resampling is performed to estimate uncertainty, entire independent sequences along with their paired results must be transferred. For humans, donors, and cell samples, appropriate grouping levels are determined by the assay device.

The report should contain the magnitude of the difference, its uncertainty and the meaning of the chosen scale. The statement “the data are not precise enough to distinguish between the null benefit and the advantage of practical interest” is more useful than the statement “the models are the same.” Equality requires its own justification. Sometimes it, as for M2 and M1I in our example, follows directly from the rules, and then statistical comparison is unnecessary.

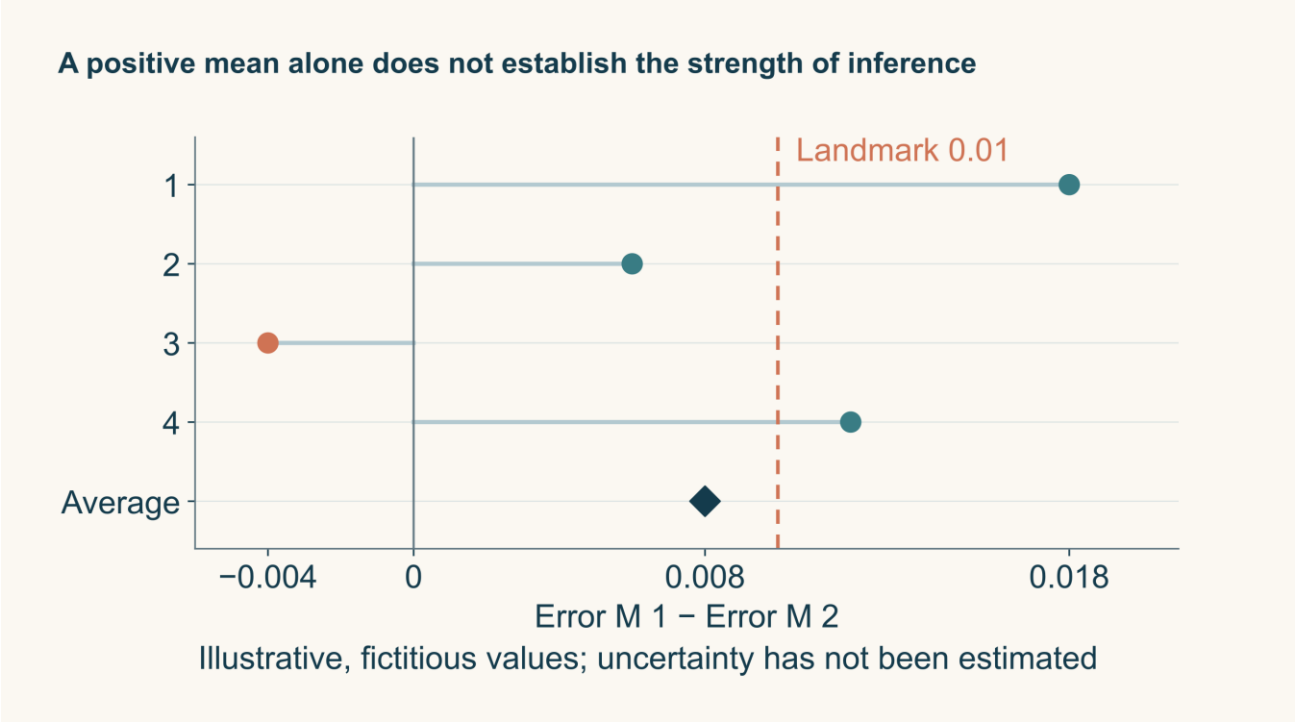

Figure 86. Illustrative pairwise differences and reference point. Four specially invented values of the chapter 35 give the average 0.008 with a conditional reference 0.01. The graph does not contain a calculated confidence interval: one cannot conclude either the equality of the models or the absence of the advantage of interest based on the sign of the mean or its position relative to the reference point.

Finding the Source of Advantage

The average of all events combines different situations. The model can especially win when the source returns, lose immediately after its change, and coincide with the competitor over long, homogeneous sections. This figure helps test the explanation for the benefit. If a separate story is important specifically when there is a conflict of stories, the corresponding condition should have special significance.

Separate the pre-planned analysis and the pattern found after viewing. Both can be meaningful. In the second case, formulate a new prediction and test it on independent sequences. You should not remove inconvenient areas and then report the remaining gain as the overall result of the original test.

Check also the absolute quality. The best of the three predictions may remain systematically wrong. For example, all candidates may attribute high resilience to a source that changes frequently in a new environment. Then the advantage of M2 does not eliminate the general defect. The next step is to examine the pattern of change in persistence itself, rather than simply increasing the number of memory entries.

Distinguish the contribution of a component from its presence

Ablation answers the question about a specific implementation in a specific task. If memory removal worsened the outcome, this shows its contribution under the same comparison conditions. It does not follow from this that without just such memory the problem is generally unsolvable. Another mechanism may store the necessary differences in a different form.

The opposite situation occurs when removing a component does not change anything. Check if it has been used. If the strategy revision condition is never satisfied, the match between the full version and the fixed version does not evaluate the usefulness of revision in situations where it actually occurs. Such a case was discussed in the engineering workshop of chapter 30. It identifies the boundary of the performed check.

To obtain a discrimination test, set an environment for which the component is expected to turn on and separately measure the frequency of activation. Then compare the result to a control that has comparable access to information and computing resources. New conditions should be called new research. The previous run does not become positive retroactively.

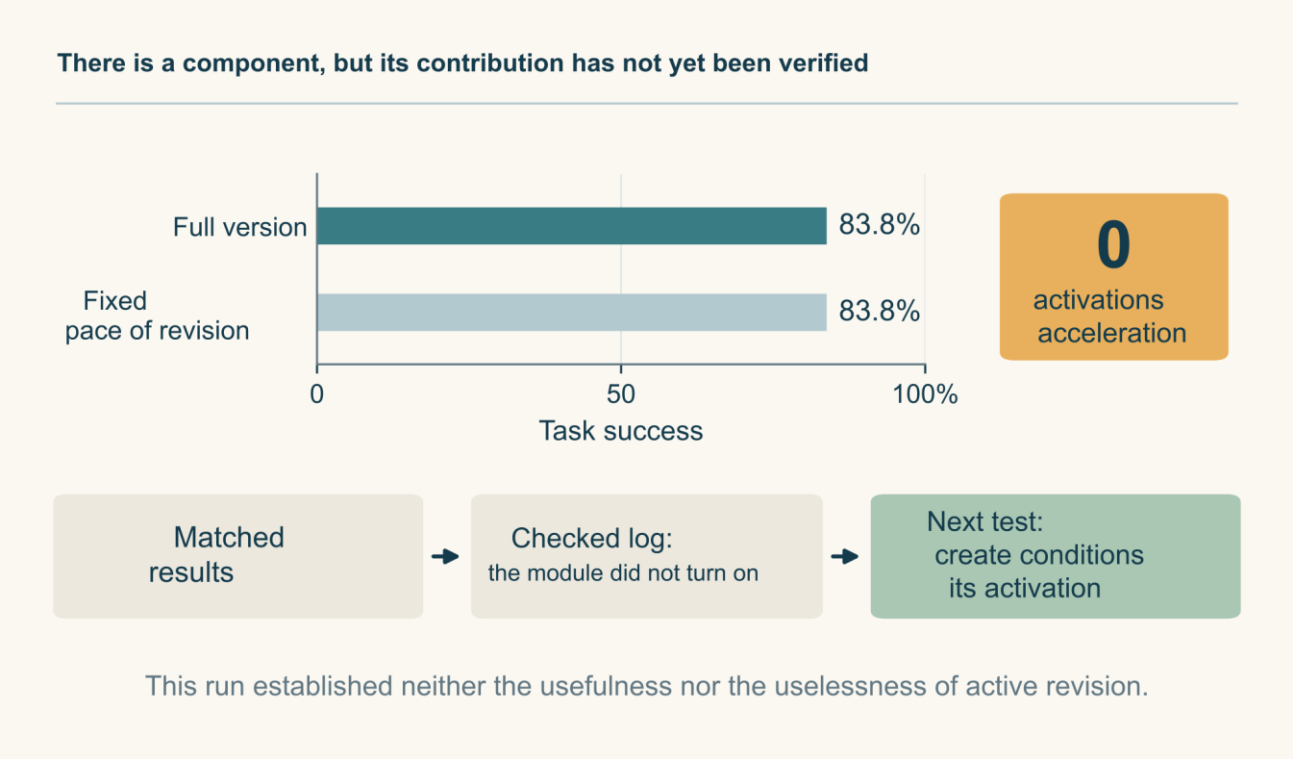

Figure 87. Honest non-discriminating result of work 8. The result of the saved program run, mechanisms_results.json, work 8: 1000 steps of one script. The full version and the fixed_revision version each have 838 successful steps (83.8%); In the full version, accelerated learning was not turned on even once. This is an exact match of a given run and not a statistical equivalence of mechanisms across all environments. The next component activation test will be a new study. Indicators do not assess qualia.

Review the mechanism openly

Let us assume that as the source absence time increases, its influence weakens more than predicted by M2. One way is to select the retention factor anew each time. A more meaningful way is to formulate a new model of time in which the record changes without reference to the source. Chapter 33 showed why these rules differ.

The new version has a risky consequence: with the same number of observations, the duration of the pause should independently influence the result. The competitive version may explain attenuation by the number of intervening events. Therefore, the next experiment differentiates the duration of the pause and the number of such events. Revision becomes scientific when it creates a new difference rather than just flattening the old curve.

Keep a table of versions: what observation motivated the change, what rule was replaced, what is now expected in the new condition. This history protects the substantive continuity of the project. The failed version may remain useful as a control or special case even if it is no longer the main explanation.

Return to the original phenomenon

Now let's imagine a clearly hypothetical result of a human study: M2 predicts the dynamics of membership assessments better than M1, but the general identity model M1 predicts it just as well. A valid conclusion is to support a separate history of causes or objects. The statement about the special mechanism of the self is not yet separated from the general recognition of the source.

The next question should concern this difference. It is possible to compare the transference of the history of a bodily source and an ordinary external object, balancing the available signs and the difficulty of discrimination. If additional parameters are entered only for the body, their bases are set in advance. Otherwise, “special self” turns into a name for additional freedom of adjustment.

The strong principle of chapter 26 requires one more step: testing the predicted changes in quality organization under targeted interventions in mutual dynamics. The success of the reporting model provides the raw material for such a project. It does not complete the experience argument. A research program advances when the next experiment distinguishes precisely those explanations that the previous one left compatible.

What a Mean Difference of 0.008 Allows Us to Conclude

Four specially invented differences M1 minus M2 are equal to 0.018; 0.006; minus 0.004; 0.012. Sum 0.032 and mean 0.008. The sign of the mean supports the

direction of advantage in this illustration, but one series has the opposite sign and the mean is below the training benchmark of 0.01. Without an assessment of uncertainty, neither an established important advantage nor the absence of such an advantage can be declared. The next action is determined by what range of effects is still compatible with the data.

What has now been established

The output is related to size, precision, and condition area. The new mechanism should generate a new discriminating forecast, and not just more conveniently describe the curve already seen.

Practice 35 Write a solution based on the result

Why do it: Make an informed decision based on one saved result.

What you will need: Preliminary protocol, Brier, paired difference and component activation. Lack of a convincing effect is not the same as proven equality.

Time reference: 30–40 minutes. Can be done intermittently; This is not a speed standard.

1. Take one completed run and reconstruct several steps: available data, preliminary forecast, open outcome, and new state.
2. Test a control with a known result. For history, this is a one-channel equality; for a component with a threshold, this is the activation log.
3. Record the magnitude of the difference, the unit of repetition, and the available estimate of uncertainty. Do not replace the missing interval with one middle sign.
4. Choose a solution: test the transfer, clarify the rule, change the non-discriminating conditions, or abandon a specific version. Name two future outcomes that lead to different decisions.

What to save: A short solution to the result with the base and the next test.

Tip: If you had to change the settings after reviewing, save the old version and name the new review a separate study.

Exploratory continuation: Disentangle the effects of pause length and number of interfering events in a new memory model. Get the forecast before collecting new data and compare with the previous version.

Analysis and answers - Appendix F, chapter 35, p. 297

Self-test

1. What does it mean that versions with and without a module match if the module is never turned on?
2. Why does the advantage of M2 over M1 not prove a special mechanism of the self?
3. How is meaningful revision different from random adjustment?

Now you can put together your own small work in which the question, mechanism, method and conclusion are agreed upon. The final chapter will offer a route from reproduction to self-change.

Chapter 36

Your Own Research

The book's route ends with research that another person can repeat and meaningfully challenge. A small mechanism is enough to begin with: its value is determined by the clarity of the issue and the strength of the test. A separate history of sources will provide a complete computational example. The next project will add the reciprocal determination of attribute and source and show what new grounds will be required for a stronger inference about experience.

From educational result to own task

The purpose of the book is to learn how to construct and test explanations of experience, and then to design systems that, on these foundations, can become candidates for subjective experience. The final work should connect design, mechanism and evidence. The scale of implementation can be small if the question is precisely defined and the result truly changes our understanding of the chosen mechanism.

Below is a completed educational computational mini-study. It shows a form of argument that can be reproduced and extended. Its numerical results follow from given rules; they are not derived from humans or cell cultures. After this, we will formulate the next project, aimed at a stronger hypothesis about quality organization and the presence of experience.

Educational study of the separate history of sources

Question. Is keeping history separate enough to cause the same new evidence to have different effects on the assessment of cause once the source is returned? Additional question: does the resulting difference require a special mechanism of the self or is it reproduced by a general memory of the identity of objects?

Method. A sequence of four observations is given: A with a positive sign, B with a negative sign, A again with a positive sign and B again with a positive sign. The source labels are available to the model and are error-free in this exercise. A positive sign has a probability of 0.75 for a common cause and 0.25 for different causes. The initial probability of the common cause is 0.5.

Before processing the feature, the previous belief is shifted to 0.5 with a retention coefficient of 0.8. The chapter update rule 25 is then applied. M0 starts from 0.5 each time. M1 carries one belief through all events. M2 stores a separate belief for

each source; the hold is applied the next time this record is accessed. M1I implements the same individual records as a regular object tracker.

This is a deterministic parsing of a given sequence. There is no estimate of the average effect in a population of sequences and no statistical sampling of participants. Each number can be checked manually or by the source_history_example() function of the engineering workshop. The output of the model refers to the probability of a common cause after the current observation, rather than the prediction of the observation itself.

Results and explanation

Step and event	M0	M1	M2	M1I
1 A positive	0.750000	0.750000	0.750000	0.750000
2 B negative	0.250000	0.437500	0.250000	0.250000
3 A positive	0.750000	0.710526	0.875000	0.875000
4 B positive	0.750000	0.858108	0.562500	0.562500

In the third step, M2 returns to story A. Its previous estimate of 0.75, after holding, turns into a preliminary probability of 0.7. Positive evidence raises it to 0.875. M1 comes to this step after negative event B, so it gives approximately 0.710526. The difference arose from the history transfer rules, although the current attribute is the same.

The fourth step is especially useful for understanding the mechanism. The overall history of M1 is now predominantly supportive, and the new positive evidence gives 127/148, approximately 0.858108. M2 returns to the previous negative history B and receives only 0.5625. Separate memory doesn't just increase confidence: it preserves different bases and can therefore lower scores relative to shared memory.

In all four events M1I coincides with M2. This equality follows from the same rules, and not from a random lack of data. Renaming the object dictionary to "model itself" does not create an additional mechanism. The computational result establishes the role of story separation and simultaneously limits its interpretation.

Controls and conclusion of the mini-study

Renaming A and B should not change the numerical sequence if their histories are transferred along with the labels. If you replace all the marks with one, M1 and M2 will coincide. If you set the retention to zero, all versions will produce 0.75 after a

positive sign and 0.25 after a negative sign. These consequences allow you to check whether the implementation actually corresponds to the stated model.

Conclusion. Separate history storage is sufficient to reproduce the described computational effect of source return. The example does not establish an advantage in any medium and does not distinguish special self-reference from general identity tracking. It provides a precise mechanism that can be tested against independent data, and a clear alternative to work with further.

The restrictions here are meaningful. An error-free mark makes the task easier; a real system must recognize the source from observations. The time rule is chosen by references, not by physical duration. The belief about the cause is not yet associated with the qualitative structure of perception. Each constraint specifies a different next hypothesis rather than invalidating an already established result.

Next Mutual Content Determination Project

Let's expand the task so that the source influences the interpretation of an ambiguous attribute, the interpretation changes the source's assessment, and the system can choose an additional check. The result will be a specific candidate for part of the organization proposed in chapter 26. First, let's set two separately changeable content properties: attribute rating and source rating. The third variable will determine the expected benefit of the test.

Before launch, we will describe targeted interventions. A forced change in source evaluation while retaining sensory input should shift the interpretation of the feature and the choice of test. Changing the representation of a feature should predictably change subsequent evaluations of the source. Breaking one feedback loop should disrupt the corresponding direction of influence while maintaining the availability of the original observation. The sizes and time windows of effects are specified by specific implementation rules.

Now comes the discriminating project. The complete system is compared to a system that has the same observations and historical records, but only one direction of influence. Separate control reproduces normal outputs without causality between components. Not only the success of the task is checked, but also the response matrix to the interventions. In this case, a fairly strong alternative controller is needed: defeating an obviously helpless recording will not explain the uniqueness of the mechanism.

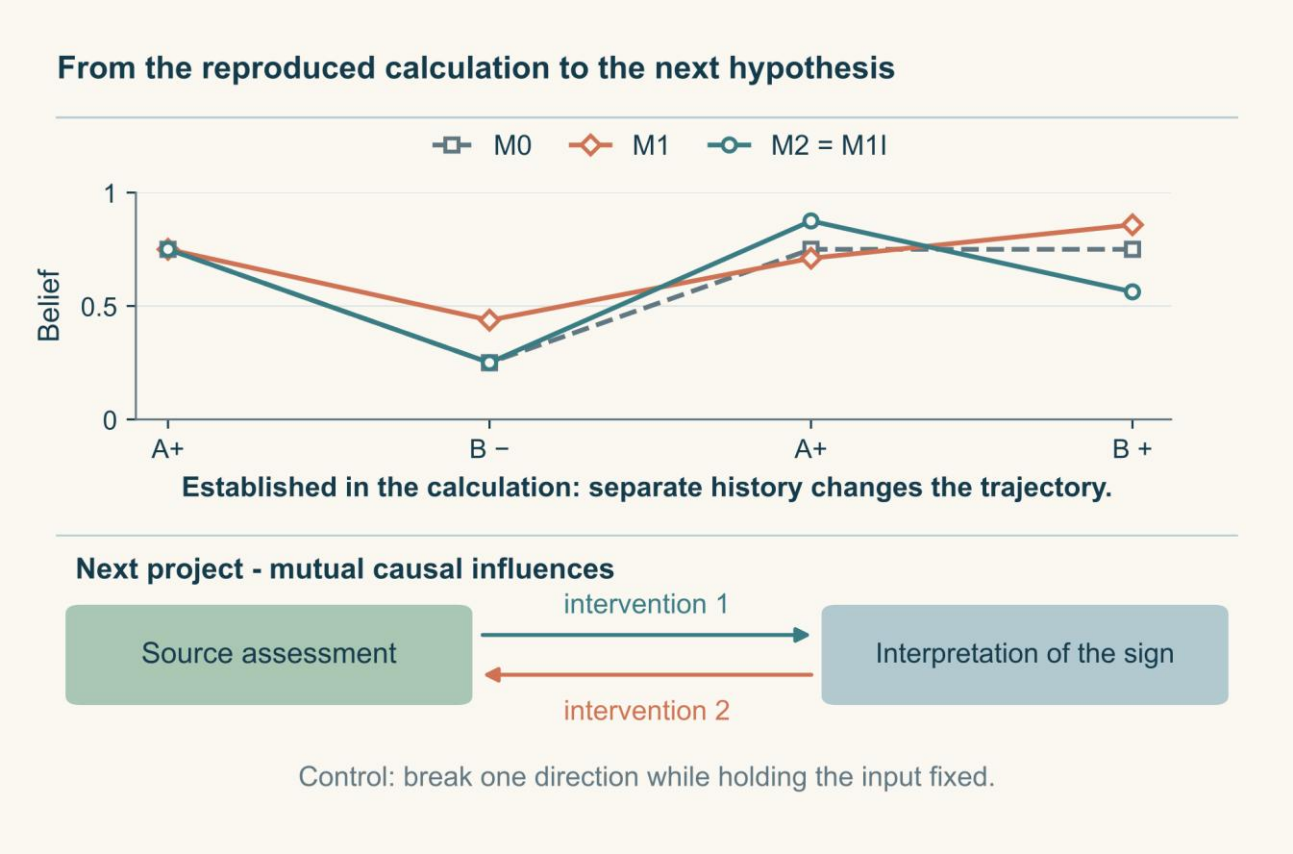

Figure 88. Completed calculation and next causal design. Above is an accurate training calculation of four events A+, B−, A+, B+. At the fourth step, M0 gives 0.75, M1 - 127/148 (approximately 0.858108), and M2 and its coinciding M1I - 0.5625. Below is a project not yet completed: interventions in source evaluation and attribute interpretation must have predetermined consequences in the opposite component. Unilateral control tests a specific causal relationship. Calculation by itself does not realize the entire strong hypothesis of Chapter 26 and does not establish the presence of experience.

How will the project approach the question of experience?

Chapter principle 26 links quality organization with sustainable relationships of distinctions and alternatives. Therefore, the future model should predict the pattern of similarities, mixtures, and content changes, not just the correct answer. For example, source interference should change some relations between ambiguous contents while maintaining prespecified other relations. This is a stronger commitment than improving overall accuracy.

Comparisons with humans will require independent data on these differences: ratings of the similarity of experiences, their dynamics, source discrimination, and control for difficulty. The proposed mechanism is compared with an explanation through general memory of objects and with an explanation through a change in decision while the experience remains unchanged. These competitors must make their own predictions before collecting supporting data.

The strong assertion of experience is retained as a hypothesis: a certain actually realized organization is sufficient for a minimal subjective process. Implementation and functional checks will verify that the proposed conditions are met. Justification of sufficiency will require connection with phenomenal data, clarification of physical implementation, and analysis of alternatives. Thus, the task of creating a subject remains a direct research goal, and each stage has its own verifiable result.

Each statement requires its own verification

Statement	Validation action	Content control
Source history influences answer	Remove or replace this particular one record	General tracker of the same story
There are two directions of influence	Intervene in p and separately in q	Two one-way circuits
The indicator improves the forecast	Test fully held-out groups	Original activity without new one sign
Organization is linked to experience	Clarify the principle and divergence consequence	Another organization with the same outputs

A benchmark comparison limits a specific conclusion, not all issues at once.

Figure 89. From approval to verification and control. Project Completion Training Matrix. Each line links the statement to a specific action and strong control. The last line preserves the independence of the theoretical argument about experience; The functional checks above do not automatically replace it.

What should be included in the final work

Prepare a coherent text with the question, rationale, mechanism, method, results and discussion. For a digital implementation, please include the executable code, environment description, parameters, stored data, and method for retrieving the tables. In the biological field, analysis of open data, a digital model and a professional experimental design with a clearly indicated boundary of the work already completed are acceptable.

Include at least one meaningful competitor, one causal or discriminative test, and one control with a known outcome. State the correspondence principle separately if you are making a statement about an experience. Indicate which of its provisions have been verified, which are assumed, and which are still underdetermined. An author's proposal requires verification of related published ideas; independent formulation does not in itself establish scientific priority.

A non-discriminating result can also become a full-fledged final work. It is important to show why it occurred and what next test will eliminate the

uncertainty. The presence of a module, its beautiful description and convincing self-report do not replace the demonstration of its role. This insight is especially useful when it allows you to eliminate unnecessary complexity and more accurately design the next candidate.

How to turn a calculation into a short article

Example question: Does the individual history change when the source is returned? The method specifies four events, reliability, retention, and memory rules. The result shows all trajectories and the coincidence of M2 with the general tracker. The discussion links the effect to the separation of entries and maintains the limitations of labels, time, and relation to experience. The next step changes one assumption and produces a new forecast. This structure allows the calculation to be reproduced and the explanatory transition to be critiqued separately.

What has now been established

We have a completed functional result and an open research goal. The next project should not blur this line: it should add a new defined obligation and a suitable test.

Practice 36 Complete your own article

Why do it: Prepare your own reproducible work with one testable mechanism change.

What you will need: Specification of the mechanism, a strong competitor, a preliminary plan and a decision on the result. The novelty of the formulation does not establish scientific priority.

Time guide: Several separate sessions after reproducing the example. Can be done intermittently; This is not a speed standard.

1. Play back the four events M0-M2 and the match with the common tracker. Save the code, parameters and raw numbers.
2. Select one change: forgetting time, source recognition errors, or the mutual influence of the feature and the source. Before launching, write down the difference from the original model.
3. Prepare a strong competitor, a discriminative test, and a control with a known answer. Check and save all results.
4. Write a paper with question, reasons, mechanism, method, result and discussion. In the discussion, separate the established fact, the remaining uncertainty, and the next question about the experience.

What to keep: Study text, executable example, data, and a clear version history of the model.

Hint: A non-discriminative result is acceptable if it is clear why it arose and what new condition would separate the explanations. Do not replace this analysis with a beautiful self-report of the system.

Exploratory continuation: For a strong hypothesis, specify interventions in both directions and infer changes in relationships between contents in advance. Comparison with phenomenal data and transfer to a new physical implementation remain separate stages of the program.

Analysis and answers - Appendix F, chapter 36, p. 297

Self-test

1. What did the above calculation of four events prove?
2. Why does a strong project need interventions in both directions?
3. Could the final work contain significant uncertainty about the experience?

Further work can follow a digital, comparative or biological branch. Keep a specific question, a complete history of changes, and the possibility of getting a result that will force you to reconsider the explanation.

Appendix A Working Dictionary

This dictionary consolidates the use of terms in the book. When reading another author, compare definitions: the same word can mean different research problems.

Qualia. The qualitative aspects of an experience, the way something is perceived or felt. The working definition does not include in advance infallibility of description, ineffability, or independence from the physical world.

Subjective experience. The experienced side of the state. The question of having experience is different from the question of what information the system discerns or reports.

Consciousness. A term with several uses. The book clarifies each time whether the presence of experience, its content, the state of wakefulness, or the availability of information for further use are discussed.

Subject. The supposed bearer of subjective experience. The definition does not require human biography, a speech interface, or membership in a pre-selected species.

Self-awareness. A family of forms of the system’s relationship to itself. Bodily identity, authorship of action, representation of one's own history, and explicit judgment of self are examined separately.

Reflection. Consideration of one's own judgments, reasons or methods of action with the possibility of evaluating and revising them. A simple numerical forecast of success does not yet realize all these possibilities.

Metacognition. Monitoring and regulation of cognitive processes. In a specific task, you need to name what is being assessed and how the assessment affects behavior.

Introspection. Referring to one's own states to notice or describe them. The nature of access and the reliability of the report require investigation.

Functional introspection. The ability of a system to use information about its internal states is being studied. The term itself does not resolve the issue of experiencing these states.

Body ownership. The experience or judgment that a particular part of the body belongs to oneself. A model must specify a measurement associated with this characteristic.

Authorship of the action. The relationship between yourself, the command and its consequences. Successful control, causal influence, and experienced authorship cannot be considered a single indicator.

Model of yourself. An idea of one's own conditions, capabilities, boundaries or history. It is necessary to clarify the content and causal role of a specific model.

Model of the world. A representation of an external situation that allows one to estimate hidden states or predict events.

Hidden state. A variable known to the environment or researcher, but not directly accessible to the agent in a given task.

Observation. Information that the system actually receives through the provided interface. The presence of a variable in the environment code does not make it an agent observation.

Conviction. In a probabilistic model, an assessment of possible states taking into account prerequisites and available information. This technical usage does not attribute experienced conviction to the program.

Prior probability. Assessment before taking into account the evidence in question. It may include the results of previous training.

Posterior probability. Evaluation after updating based on new evidence within the selected model.

Likelihood. The probability of the observed data given the assumed state or model. It must be distinguished from the probability of the hypothesis itself after observation.

Uncertainty. Limited knowledge of a condition, parameter, or explanation. Different sources of uncertainty may require different actions.

Calibration. Correspondence of probabilistic forecasts to the frequency of outcomes in certain groups of cases. Its assessment depends on the data and the method of grouping.

Brier score. The mean square of the difference between a probabilistic prediction of a binary outcome and an outcome written as zero or one. Characterizes the quality of the forecast, not just the calibration.

Usefulness. A specified way to combine results and costs. In the first paper, these are successes minus the cost of additional observations.

The value of information. Expected improvement of the solution due to additional information, taking into account the goal and the adopted model. Reducing uncertainty is not always equally beneficial.

Feedback. Returning information about the result to further operation of the system. It is necessary to indicate what information is returned and with what delay.

Recurrent processing. Processing with influence return: the result changes the operation of the elements that already participated in its formation. The specific computational and physical organization must be described separately.

Workspace. Organizing the selection and sharing of content across multiple processes. A simple table of variables may model individual functions, but not automatically the entire scientific theory.

Ablation. Removing or disabling a component to test its contribution to a specific implementation and task.

Causal intervention. A directed change in a variable or relationship designed to test its influence. Spillover changes and the accessibility of the intervention to other processes are included in the analysis.

Fictitious intervention. A control procedure that preserves the significant variable under study while reproducing other aspects of the effect.

Correlates. A sign that is statistically related to the phenomenon being studied. Connection in itself does not establish necessity or sufficiency.

Necessary condition. A condition without which the expected phenomenon does not occur in a specified area. The lack of effect after an intervention requires testing the specificity of the intervention itself.

Sufficient condition. A condition whose presence should ensure a phenomenon under explicitly specified prerequisites. It may not be the only possible way to obtain the phenomenon.

Invariant. A property that is preserved under a certain class of transformations. The transformation class and measurement method must be named in advance.

Identifiability. Ability to distinguish models or parameters based on available data. Different names for mathematically identical algorithms do not create identifiability.

Overfitting. Adjustment to the characteristics of the data used, which does not provide the expected quality in new cases.

Control sample. Data used for specified verification after model development or evaluation. Re-tuning them changes their research role.

Pre-registration. A description of the question, comparisons, indicators and analysis saved until the appropriate stage. It makes changes to the plan noticeable, but does not guarantee the quality of the original idea.

Unit of observation. The object or incident to which the measurement relates. It should be distinguished from an independent unit of experimental repetition.

Pseudoreplication. Mistakenly treating dependent measures as independent replicates. Many episodes of one sample do not replace many independent samples.

Bootstrap. Obtaining repeated samples from existing data to estimate the variability of statistics. The reselection method must be consistent with the dependency structure of the observations.

Preprint. An accessible version of a research text prior to or beyond completed journal publication. The status and version must be checked separately.

Neural culture. Experimental system with grown nerve cells. Its organization depends on the specific model and conditions.

Organoid. A three-dimensional cellular model that reproduces some features of tissue organization or development. An organoid should not automatically be considered a whole organ or organism.

Hybrid system. A system in which biological and technical components interact causally. The analysis must show the contribution of each component and the selected boundary of the whole.

Phenomenal conclusion. A statement about the existence or nature of subjective experience made on the basis of data and theoretical premises.

The principle of correspondence. A proposed relationship between the description of the system's organization and the description of the experience. In the author's hypothesis, it requires an independent argument and a definition of the scope.

Likelihood: The likelihood of a given data given the version chosen. It is not equal to the probability of the version itself after the data; the latter requires prior probability and normalization.

Conditional independence: An assumption under which the probability of a new event for a given cause does not depend on already known events. Copying a signal usually violates this assumption.

Brier error: The mean square of the difference between the probabilistic prediction of a binary observation and its outcome is zero or one. The forecast is recorded before the observed observation occurs.

Model recovery: Testing the choice of mechanism on artificial data where the generation mechanism is known. Reconstruction errors show the limitations of the discrimination procedure.

Deferred group: All records of a selected organoid or batch that are excluded entirely from training for a specific transfer test.

STTC: Spike time tiling coefficient - an indicator of the temporal coincidence of discharges. The training analysis uses the average absolute value of the finite off-diagonal elements of the matrix at the window 15 ms. This is not a proven causal relationship.

Sufficient condition: If the condition is satisfied, the intended conclusion follows within the limits of the accepted statement. Failure to satisfy such a condition does not in itself exclude other grounds for the same conclusion.

Appendix B Launch and arrangement of the workshop

Preparation

Unpack qualia_mira_practice.zip into a separate directory. Inside there are programs, instructions and saved results. Start with the README.md file. Don't run Python files directly from the archive viewer: they need adjacent files. For work 2–8 agent_lab.py and mechanisms_lab.py must be nearby. It is convenient to save new results under different names so that the original examples remain available for comparison.

Open a terminal in this directory. On Windows, this can be done through the Explorer menu or by using the cd command. On macOS and Linux, use a regular terminal. Check your Python version: 3.10 or later is required. If Python is not already installed, use the official python.org distribution for your system. The name of the command depends on the installation.

```
python3 --version
python3 agent_lab.py --self-test
python3 mechanisms_lab.py --self-test
```

On Windows, the typical form of the first command is py -3 --version. Similarly, replace python3 in the remaining commands. If the system reports that the command was not found, this is an error in starting the interpreter, and not the result of the experiment. If agent_lab.py is not found, check the current directory. If mechanisms_lab.py cannot import agent_lab, check that both files are unpacked together.

For the first acquaintance, you can complete all eight works and pair analysis with a single command. It creates a separate results_... directory with the results and the run_manifest.json file, which contains the parameters and checksums of the programs.

```
python3 run_practice.py
```

One job can be selected with the --lab parameter, for example python3 run_practice.py --lab 6. On Windows, use py -3 instead of python3. General run of tests without saving a new experiment: python3 run_practice.py --self-test.

How to read result files

JSON stores data in the form of named fields and lists. You don't need to master the entire format at once. In lab01_results.json, the summary section contains the general results of the modes, and per_run contains the results of individual

scenarios. In mechanisms_results.json, the works are located inside labs under numbers from 2 to 8. Detailed explanations of the fields are included in the archive instructions and chapters 28–30.

The success_rate and utility_per_step fields indicate the success rate and utility per step. The value 0.8 for the success rate corresponds to eighty percent. The queries field contains the number of additional queries, and brier_score is an indicator of the quality of the probabilistic forecast. When comparing tables, always check that the number of steps, parameters, and averaging method are the same.

First reproduce the original result. Then create a new file with one parameter changed. Keep the command text and the preliminary expectation nearby. After checking on the original seed, use another one. Random grain ensures repeatability of a particular scenario; it does not make a single scenario a sufficient study.

Eight works

Work	Program	Main material
1	agent_lab.py	Memory of the world, model of action, value of observation
2	mechanisms_lab.py --lab 2	Geometry of representations and invariants
3	mechanisms_lab.py --lab 3	Reprocessing and detailed control
4	mechanisms_lab.py --lab 4	Limited selection and general access
5	mechanisms_lab.py --lab 5	M0, M1, M2, common tracker and source delay
6	mechanisms_lab.py --lab 6	Designed internal access and interventions
7	mechanisms_lab.py --lab 7	Choosing a strategy based on past mistakes
8	mechanisms_lab.py --lab 8	Composite candidate and component contribution control

All operations 2–8 can be performed in one run without the --lab parameter. The command below will create a new result file. It does not send anything to an external service and does not change the code.

```
python3 mechanisms_lab.py --json my_mechanisms.json
```

The programs are designed to study small mechanisms. Their detailed structure is described in the book. Combining function names into one system does not replace separate testing of how the components interact. The eighth paper specifically

includes comparisons in which the added mechanism does not provide an advantage in the base condition.

Additional step-by-step models

The teaching_lab.py file complements the original eight papers. It uses the standard Python library and when analyzing chapter 31 reads the nearby bio_features.csv. Execute the commands from the practice folder of the unpacked kit. The python command indicates the available Python interpreter 3; on some systems it runs as python3 or py -3. Basic tests and training models do not require the installation of external packages.

```
python teaching_lab.py --self-test
python teaching_lab.py --chapter 13
python teaching_lab.py --chapter 25
python teaching_lab.py --chapter 31
python teaching_lab.py --chapter 33
python teaching_lab.py --chapter 34
```

Each run saves JSON with rules or numeric results. For chapter 31, a ready-made table allows you to repeat the main analysis without loading a large array. The extract_bio_features.py script takes the official file [47], checks its checksum and repeats the feature extraction with obvious exceptions. This particular step requires NumPy; details and attribution are in the workshop README.

Appendix C Paired comparison of results

In the first work, all four modes use common random scenarios. Therefore, the natural unit of comparison is the difference between the performance of two regimes within one scenario. If we first combine all the steps and declare them to be independent iterations, we get an unjustifiably strong impression of precision: adjacent steps are related by the goal state and the agent's learning.

The analyze_lab01.py program calculates the utility differences between the full regime and each control regime for thirty scenarios. It then resamples the integer differences four thousand times and constructs a percentile interval. The method describes the variability given the assumed repetition structure; exact ninety-five percent coverage for any sample is not guaranteed.

```
python3 analyze_lab01.py lab01_results.json --json paired_analysis.json
```

In the retained analysis, the average utility advantage of the full mode per step is 0.08516 over the memoryless version, 0.37772 over the fixed action model, and 0.08280 over the no-request version. The corresponding intervals are approximately 0.07975 to 0.09080, 0.37133 to 0.38434 and 0.07583 to 0.08997. These values are obtained from the attached data and analysis algorithm.

Intervals refer to specific comparisons, metrics, and the built environment. Multiple separate intervals do not automatically result in a joint ninety-five percent guarantee. If, after viewing many indicators, you choose only the favorable one, you need to take into account the search nature of such a decision. In a full-fledged study, the main criterion and family of comparisons are described in advance.

For independent work, choose utility as the main indicator, and use accuracy and cost to explain the mechanism of differences. First, state the expected direction of the effect. Then look at the differences between individual scenarios: is the result consistent, are there any unusual cases and are they associated with a specific sequence of events? Do not delete an inconvenient script without a predetermined reason.

If the mean difference is small, various explanations are possible: the mechanisms are truly similar in the chosen task, the data are insufficient, or the measure does not reflect the expected property well. A wider interval does not automatically select between them. The next experiment should be related to a specific cause of uncertainty: a new task, additional independent repetitions, or a more appropriate measurement.

Appendix D Forms for independent work

Recording the Experienced Episode

Indicate the situation and moment of the event. Note when the recording was made: during the episode, immediately after, or later. Describe what you noticed and keep unclear areas separately. Then write down your explanation and at least one alternative. Finish by asking what observation would help clarify the conclusion. The margins can be filled in with free prose; the absence of a noticed detail should be left as an absence, and not completed according to the model.

Research paper card

Write down the authors, year, title and link. State your research question in one sentence. Identify participants or samples, conditions, measurements, and comparisons. Separately state the result obtained and the conclusion of the authors. Then write down the assumption that connects them, the possible alternative, and the next discriminative test. If the article is theoretical, instead of sampling and measuring, describe the premises and implications.

Model card

Name the phenomenon being explained and its scope. List available inputs, hidden states, memory, and update rule. Explain how the model produces the observed response. Specify the parameters and how they are determined. Conclude with a description of the closest competitor and conditions where forecasts differ. For the experiential hypothesis, add the correspondence principle and an argument in its favor.

Recording a Computational Experiment

Save the question, code version, command, parameters, seed, base metric, and expected diff. Once run, add the result, the data path, and an explanation for the deviations. Note which decisions were made before looking at the data and which ones after. The next test run must be separated from the setup phase.

Final Project Specification

State the subject of the explanation, the proposed mechanism, method of implementation, available data, intervention design, and comparable alternatives. Describe a positive criterion for success and an outcome that would require a change in hypothesis. Then list the actual checks performed. Finally, separately

formulate the functional result, the theoretical interpretation, and the basis for a possible conclusion about the experiment. These paragraphs should be consistent while maintaining the distinction of their objectives.

Appendix E Index of Illustrations

The pictures are located next to the explanations. Captions indicate the status of the image and the origin of the numbers. Schemes reveal selected relationships; subject illustrations help to present an episode or object under study. The number in the index leads to the picture.

1. Three areas of description of one episode 13
2. Recording time as part of surveillance 14
3. Comparison of surfaces in one episode 19
4. Names of materials and similarities of touch 20
5. From pairwise estimates to a similarity map 21
6. Accessory without movement and control of external object 27
7. Four attitudes towards yourself in one episode 29
8. Confidence with equal success 36
9. Checking logical inference with counterexample 281
10. Causal influence and common cause 41
11. Three research roles in the design of the ..... 42 lamp
12. Participant and Observer in Study 46
13. Coordination of different time data 48
14. Question time as a condition of procedure 50
15. Material distribution and order conditions 54
16. Total budget and preparation time distribution 55
17. The magnitude of the effect and the accuracy of its assessment 60
18. Responses and participants as different data levels 62
19. Dynamics of accumulation and discharge 67
20. Input allocation and state saving disabling 69
21. Two ways to use context 74
22. Selective violation of channel usage 75
23. Hand tool and observed result 78
24. Sequential update of causal versions 79
25. Distinguishing causes by changing command 82
26. Information before and after feedback 86
27. Predictions and error of two models 88
28. Recurrence and global access 94
29. Independent shutdown of two mechanisms 97
30. Two different objects explanation 101
31. External log and internal use 103
32. From preliminary estimate to updated probability 107
33. Separate models of the world and actions 108
34. Report and measurement limits 114
35. Why lack of significance is not exactly zero 116
36. Three trajectories of one teaching mechanism 120
37. Averaging and dynamics of a single episode 122
38. Development in everyday interaction 125
39. Decomposition of the visual self-recognition problem 127
40. Development profile without common ladder 129
41. The environment specifies different behavior tasks 132
42. Memory compromise instead of complexity ladder 133
43. Two histories of one distribution of the characteristic 134

44. Explanation of presence and explanation of quality 139
45. From device to explanation of experience 142
46. Different organisms and different research questions 145
47. Matrix of questions and evidence gaps 148
48. Comparable tasks with different external responses 149
49. Dividing the number of events and the relationship between them 155
50. Habituation and exhaustion differ by transfer 156
51. Separate audit of information and self-description 159
52. Artificial system in a physical environment 160
53. The carrier does not define the study boundary 162
54. Commitment card model 167
55. A fair comparison of computing competitors 168
56. Normalization of one update 176
57. Return to the history of the source 174
58. Renaming and changing the causal structure 181
59. From causal organization to the hypothesis of experience 184
60. Four conditions of the distinguishing project 188
61. From possible outcome to discrimination test 190
62. Two circuits of one agent 195
63. Conditional value of additional observation 197
64. Success and usefulness answer different questions 199
65. Uncertainty of differences between modes 200
66. What is saved when converting the ..... 203 view
67. Recurrence and comparable alternatives 205
68. Content delivery and direct access control 207
69. Access check after late instruction 208
70. Whose history is used when updating 212
71. History helps to recognize a delayed source 213
72. Internal access is checked by intervention 214
73. Causal order of functional revision 215
74. The revision causes a new decision 216
75. Assembly of several mechanisms into one system 217
76. The presence of a module does not yet indicate its contribution 218
77. Biological component and instrumentation 221
78. Levels of accounting in biological data 225
79. Temporal and group separation of biological data 227
80. Three different questions about the internal state of 232
81. Candidate Screening Profile 234
82. Two clearly different memory timing models 239
83. Limit of available information and ambiguity of parameters 241
84. Data structure of the preliminary protocol 247
85. Preliminary plan and subsequent opening 250
86. Illustrative pairwise differences and reference point 254
87. Honest non-discriminating work result 8 255
88. Completed calculation and next causal project 262
89. From approval to verification and control 263

Appendix F Reviews of practices and self-tests

Come back to the analysis after your own attempt. For observations, evaluate the clarity of the recording and the distinction between inference and assumption. For calculations, check the original rules and intermediate steps. Research continuations allow for several informed decisions. The self-test retains questions from the chapter; here are the answers.

Chapter 1 Analysis of practice

There is no single list of sensations here. The work is done if the external fact, description and explanation can be distinguished from the record. The greater number of details in the second attempt does not yet establish the enrichment of the experience itself: attention, memory and vocabulary could change. A good remaining question is naming the alternatives and the next comparison.

Self-test answers

1. The phrase reports a change in sensation and explains it by a physical change in the object. Reasons are needed for both the heating claim and the causal connection. One experienced change allows for several explanations.
2. Yes, if we are talking about how a specific episode was experienced. This message must be distinguished from a stable preference or general judgment about the quality of a thing. The decision is determined by the meaning of the statement.
3. Differences between records have been established. It may reflect changes in experience, attention, memory, or language. Further comparisons will be required to infer the cause. The sheer number of details is not a measure of the amount of consciousness.

Return to chapter 1

Chapter 2 Analysis of practice

The correspondence of the map to the original relationships is checked. If the numbers were given only the order, there is no need to require an exact coincidence of all distances. An unstable answer cannot be silently corrected for the sake of beautiful geometry. In the exploratory continuation, the usefulness of the model is assessed based on new, pre-predicted relationships.

Self-test answers

1. The name can unite cases by material, purpose or familiar category. Descriptions of experiences and pairwise comparisons define other relationships. It is necessary to clarify on what basis the grouping is made.
2. Its meaning depends on the instructions. It can only indicate a place in the ranking order. A statement about an exact ratio of quantities, such as “twice as strong,” requires a separate basis.
3. You should determine in advance which relationships or new cases the model will predict better. Then use the data that was not involved in the selection of coordinates. Improving the drawing on the same four points does not yet show the ability to explain something new.
Return to chapter 2

Chapter 3 Analysis of practice

The immobility of the hand does not negate the ownership, and the control of the pointer does not make it part of the body. The past authorship of a note is not identical to the current action. It is not necessary to have the same experiences in different people; it is required to maintain the differences between the issues and not to deduce the internal attitude solely from the external event.

Self-test answers

1. Control refers to the connection between the command and the outcome. Belonging refers to whether an object is given as part of one's own body. These relationships can be combined, but one does not determine the other.
2. The actual authorship is determined by the causal chain of events. The sense of authorship is studied as an experienced attitude towards action and through corresponding reports. They may diverge, and this discrepancy itself requires explanation.
3. A name is an element of self-description. It does not explain spatial perspective, bodily identity, experienced action, or the testing of one's own judgments. A utility model must indicate which of these problems it solves and by what mechanism.
Return to chapter 3

Chapter 4 Analysis of practice

Answers: erasers; the presence of a large green card is not necessary; Peter came last. In the second case, all the green ones can be small round ones, and the big round ones can have a different color. This is a counterexample that satisfies the conditions and refutes the necessity of the conclusion. The task shows the relationship between a prediction and a subsequent test, rather than measuring the general level of self-awareness.

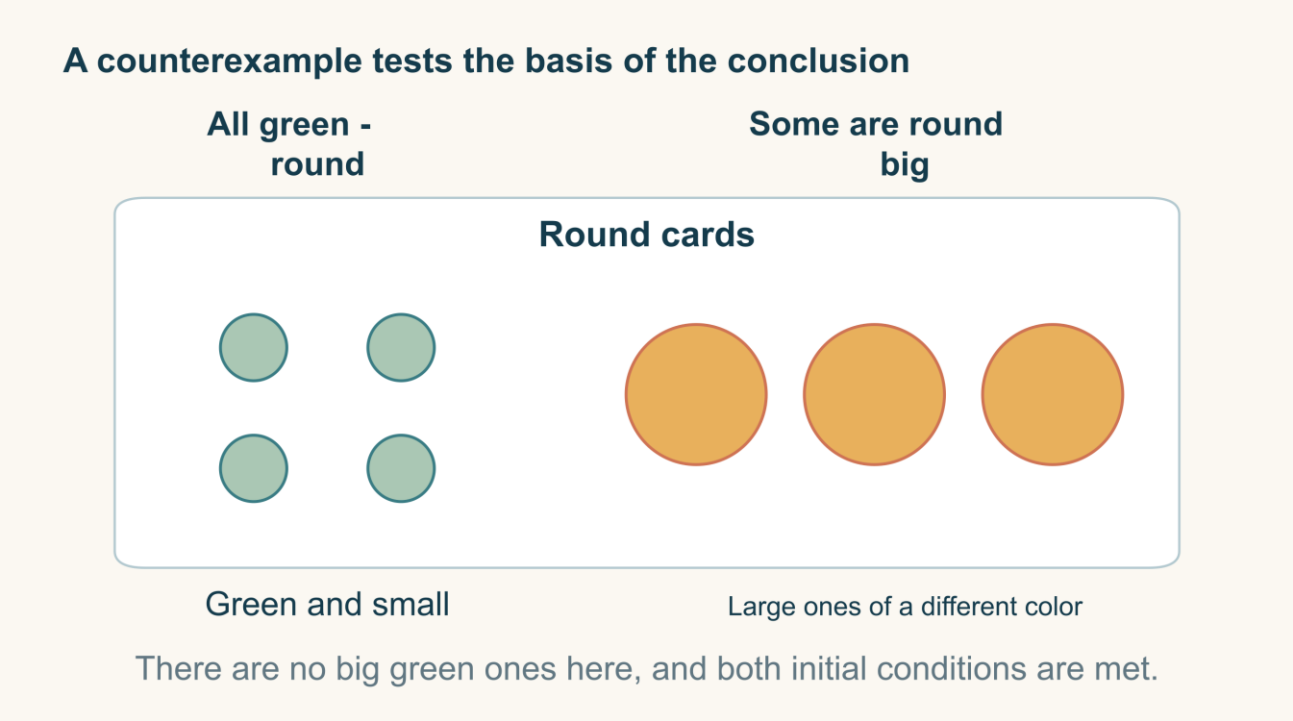

Figure 9. Checking logical inference with a counterexample. All the green cards in the diagram are round, and some of the round cards are large. However, there are no big green ones: the initial conditions do not guarantee such a conclusion, and the counterexample shows a specific gap in the reasoning.

Self-test answers

1. Feedback adds information and allows the construction of a new way of explaining. It is necessary to distinguish between initially noticed steps, later memory and reconstruction taking into account the correct answer.
2. Yes. One can attribute equal confidence to almost all answers. Overall success and the ability to direct the check to probable errors will then be different characteristics.
3. State the type of task, the observed error, and the intended process, then suggest a specific change to the procedure. After this, compare the results on new comparable tasks, taking into account possible alternative reasons for the improvement.
Return to chapter 4

Chapter 5 Analysis of practice

Comparing the diary with an equal pause begins to separate the proposed versions. It still requires measurement, comparable tasks and order control. If the chosen metric only reflects an increase in description length, the conclusion about improved discrimination is not yet assured. Good work preserves this limitation and offers the following comparison.

Self-test answers

1. A mechanism shows how the parts interact to produce an outcome, including the conditions and sequence of changes. The parts list does not explain how they work together.
2. A metric can affect, depend on, or change with a report due to another process. Discrimination requires additional data and appropriate interventions.
3. What is needed is a well-founded bridge between the realized function and the experience: a theory of correspondence, its scope, and tests of alternatives. Success in recognition itself establishes the completion of the task, and further inference requires an additional argument.

Return to chapter 5

Chapter 6 Analysis of practice

Checking an item determines its external properties. The original report refers to its procedure and is not rewritten by later knowledge. A specification is useful if another person can repeat the instructions and determine exactly when the information became available. Consensus across multiple indicators requires testing common causes rather than counting votes.

Self-test answers

1. Lack of recall may occur after an episode or reflect the difficulty of the issue. It is necessary to distinguish between experiencing, storing information and the ability to communicate it at the moment.
2. It shows the successful use of information in a specific task. Conclusions about clarity, completeness of experience and other properties require additional data and a reasonable connection with them.
3. All three indicators may depend on the overall process, such as response preparation. Consensus becomes more meaningful when such common causes are tested and it is shown which alternative explanations are ruled out by a combination of methods.

Return to chapter 6

Chapter 7 Analysis of practice

The long entry may be due to dictionary, format training, and timing. To draw a conclusion about discrimination, a corresponding new result is needed. The four distributions help to separate the material, order and method of work. If all outcomes are declared as confirmation, the hypothesis does not yet specify a testable difference.

Self-test answers

1. In one condition, a person can read all the time, and in another, spend part of the time on the diagram. The comparison makes some sense given the overall budget, but the internal operations remain different.
2. It breaks the systematic choice of condition based on the known characteristics of the participants and sets the distribution procedure. However, in a small group, the composition of conditions may accidentally be different.
3. A new hypothesis arose taking into account the patterns already seen. For independent verification, new data or a previously separated part of data that was not used in its formulation is needed.
Return to chapter 7

Chapter 8 Analysis of practice

In the fourth repetition, the difference is minus twenty points, in the sixth, plus twenty. According to the conditions of the educational generator, the modes are the same, but individual assessments fluctuate. Inaccuracy may require new independent data; precise exclusion of an important effect limits the hypothesis; The dependence on the material requires a new check of the conditions. The number of rows does not replace the number of independent study units.

Self-test answers

1. Percentage points are the difference between the shares themselves. Relative growth is calculated relative to the baseline: ten divided by seventy, resulting in approximately fourteen percent.
2. The responses of the general model depend on certain parameters and training history. They help evaluate its behavior, but do not replace information about variability between independently created systems.
3. Yes, if the measurement is sufficiently accurate to exclude a priori significant improvement under specified conditions. With a wide range of compatible effects, the result remains uncertain, and the absence of a convincing signal does not prove the absence of an effect.
Return to chapter 8

Chapter 9 Analysis of practice

For A, the impulse appears at the third step. B does not reach the threshold; the final state is approximately 0.921754. When storing zero, each non-zero input gives only 0.6, so there are no pulses in both sequences. What is explained is the dynamics of the chosen mechanism, not a universal property of neurons or consciousness.

Self-test answers

1. The rule uses not only the current input, but also the saved state. Different stories change the distance to the threshold.
2. The state describes the current dynamics under these rules. Learning changes the very parameters of these rules and thereby the behavior in future sequences.
3. It distinguishes between an explanation in terms of a total sum and an explanation in terms of temporary accumulation with leakage. The difference in outputs supports the role of the temporary structure in a given model.
Return to chapter 9

Chapter 10 Analysis of practice

The direct rule and reprocessing may produce the same category. Coincidence itself does not select the mechanism. Two identical cards with different hidden labels show information ambiguity. Additional inference requires certain rules, equal available information, and a condition under which the answers diverge.

Self-test answers

1. The report depends on the question format, memory, timing and message sequence. To draw conclusions about experience, a reasonable connection with these processes is needed.
2. Context can be used by different architectures. One needs to compare models with comparable inputs and resources under conditions where their predictions differ.
3. To whom or what exactly is information available, for what operation and at what moment. The ability to categorize, report, remember, and check an answer cannot automatically be considered the same ability.
Return to chapter 10

Chapter 11 Analysis of practice

History allows you to compare the effect with the previous command. For the selected delayed series, the connection is restored with a lag of two. If two mechanisms produce the same observations in all steps taken, there is no current data to choose from. The hidden column is needed by the evaluator, but should not participate in the calculations of the model under study.

Self-test answers

1. A common cause may be signals from external objects. Bodily belonging requires additional relations to the representation of the body and data about the corresponding experience.
2. New interventions on commands and impact comparisons. If the observations in two versions are the same, passive coincidence alone is not enough to confidently distinguish them.

3. The improvement may be due to overall memory capacity. To distinguish an organization by sources, you need to compare it with suitable alternatives and check the transfer to new conditions.
Return to chapter 11

Chapter 12 Analysis of practice

Primary forecasts are approximately 0.500; 0.667; 0.750; 0.600; 0.667; 0.571. Their mean square error is about 0.309, for the constant model 0.250. This is the outcome of one short episode. When continuing, you need to start with the saved counters, and not repeat the original smoothing. It is the quality of the forecast that is tested, not the depth of the confidence experienced.

Self-test answers

1. It uses information that was previously missing. The forecast being tested must be fixed before feedback, otherwise a hint from the future arises.
2. Individual tasks may vary in difficulty and features available. The general history estimates the average reliability of the selected class, and the specific case requires additional information.
3. The update follows the specified rule. Explicit reflection considers the basis, applicability or limitations of this rule and may lead to the proposal and testing of another way of assessing.
Return to chapter 12

Chapter 13 Analysis of practice

The complete system gives 10 out of 10 for both indicators. Without return, recognition and memory give 7 out of 10. Without transmission, recognition stores 10 out of 10, memory gives 4 out of 10; when both indicators are disabled, 7 and 4, respectively. A direct competitor gives 10 out of 10. These numbers come from the device and are not derived from a human study.

```
Reproduction: python teaching_lab.py --chapter 13
```

Self-test answers

1. Why aren't global propagation and recurrence strictly opposed?
Recurrence describes the return of influence, and globality describes the accessibility of content to different processes. The general exchange system can use feedback links. It is necessary to distinguish between supposed sufficient conditions and functions of specific interactions.

2. Does the disappearance of the report after the intervention prove the disappearance of the experience?

It doesn't prove it in itself. The intervention may have impaired memory, response selection, or instruction following. The conclusion is strengthened by independent indicators and monitoring of the preservation of these capabilities, although the problem of measurement does not completely disappear.

3. Why compare mechanisms with the same computing budget?
Otherwise, the advantage may be mistakenly attributed to a specific architecture, although one system simply received more processing steps. Control helps to determine whether the organization of interactions is necessary for the observed result.

Return to chapter 13

Chapter 14 Analysis of practice

Success here means sharing responsibilities, not picking a winner. The specification should allow you to agree with the device and separately argue about the connection with the experience. Given the same predictions, there is not enough data to make a choice; in the absence of the necessary formalism, one cannot pass off an approximate scheme as a calculation of the criterion of the theory.

Self-test answers

1. Why is a high correlation between parts not sufficient to infer their mutual causal influence?
Both parts may depend on a third source. Causal inference requires additional premises, interventions, or a specifically justified dependency model.

2. Is reflection equal to any higher order representation?
No. Expanded reflection includes examining one's own reasons, comparing explanations and revising strategies. Representing the reliability of a current signal can serve a much narrower function.

3. Can a theory be mathematically precise and empirically controversial?
Yes. Precision allows consequences to be rigorously deduced from assumptions. Confirming that these assumptions and implications describe consciousness requires independent evidence and comparison with alternative explanations.

Return to chapter 14

Chapter 15 Analysis of practice

For the signal on the right, the new probability is about 0.857. For the signal on the left, it is equal to 0.12 from the sum of 0.12 and 0.32, approximately 0.273. With direct communication, the command coincides with the side of the target,

with reverse communication, it is the opposite. The mere coincidence of expectations does not prove experienced authorship; a computable causal model is tested here.

```
Reproduction: python teaching_lab.py --chapter 15
```

Self-test answers

1. Is every prediction a conscious expectation?
No. A prediction is a calculated distribution of possible signals. The presence of such a calculation does not in itself establish its experience.

2. How does a model of action differ from a model of the world?
The first describes the connection of one’s own commands with the consequences, the second describes the state and changes in the surrounding situation. Their mistakes require different corrections, even if outwardly both lead to unsuccessful action.

3. Why should the absence of a signal be distinguished from a negative signal?
A non-functioning sensor does not indicate that a target is missing. This confusion creates unfounded confidence. It is necessary to explicitly model the omission of an observation and the possible change in the world during the omission.

Return to chapter 15

Chapter 16 Analysis of practice

The correct specification may contain unknown fields. By canceling a report, one cannot automatically conclude that all related processes are absent. Decoding shows the information available for analysis; necessity and sufficiency require other grounds. If the following prediction is common to two theories, it is useful for the mechanism, but does not yet choose between the theories.

Self-test answers

1. Why shouldn't a major study finally settle the theory debate?
Its scale increases the information content of certain measurements, but does not automatically expand the scope of verified statements. The chosen predictions, objectives, metrics, and sensitivity of the analysis matter.

2. Is the no-report condition a direct measurement of experience?
No. It reduces some of the response requirements, but requires justification for how perception is determined. Indirect indicators also have prerequisites and limitations.

3. What changes confidence in a model more—the expected outcome or the discriminative prediction?

Particularly informative is the result, which is probable with one model and unlikely with a serious competitor. Concordance with a general expectation is useful, but less helpful in selecting an explanation.

Return to chapter 16

Chapter 17 Analysis of practice

Complete trajectory in steps 0–4: 0; 0.4; 0.6; 0.6; 0.6. Without maintenance after 0.6, the results are 0.3 and 0.15. Without the first input, we get 0; 0; 0.4; 0.2; 0.1. The status read mark does not change the dynamics unless specifically included in the rule. All numbers refer to the learning mechanism.

Self-test answers

1. Why is dynamic emergence not the same as evolutionary origin?
The first describes state transitions of a running system. The second concerns the history of the appearance and changes in its mechanisms between generations. They require different data and explain different time scales.

2. Does sustained activity prove the presence of a conscious episode?
No. Stability explains the persistence of a state. It is necessary to further explain the content, its causal role and the basis of the connection with the experience.

3. What does the discrepancy between neural and behavioral indicators mean?
It limits the simple identification of indicators and requires testing of several explanations: the sensitivity of the measurement, the role of the process in the problem, and the possibility of compensation. The discrepancy itself is useful data.

Return to chapter 17

Chapter 18 Analysis of practice

For example, failure to touch a mark may depend on an unnoticed detail or on the difficulty of the corresponding movement. Prominence and action controls separate these versions. Requesting help also requires social accessibility of help. The conclusion of each achievement is not the same as dating the emergence of all forms of consciousness.

Self-test answers

1. Why can’t the age of successful self-recognition be considered the age of the emergence of consciousness?
Self-recognition requires a certain set of visual, motor and cognitive capabilities. Conscious experience and other forms of self-reference are not identical to the performance of this particular task.

2. How does asking for help differ from explicit reflection?
The query can be governed by uncertainty assessment without verbal analysis. Explicit reflection presupposes that a person makes his own reasons the subject of consideration, compares them and is able to purposefully revise them.

3. What does the absence of the expected indicator for the younger group mean?
It allows for several explanations: lack of function, different implementation, additional difficulty of the task, or insufficient measurement sensitivity. Control tasks and observation of changes over time help to choose an explanation.

Return to chapter 18

Chapter 19 Analysis of practice

Both learning stories are compatible with the given leaves. The benefits of memory are relative to variability and costs. Even a confident reconstruction of the origin of M does not establish experience without a valid connection between M and experience. A good map stores all these dependencies and indicates what data is needed next.

Self-test answers

1. Does the adaptive benefit of a mechanism prove the presence of experience?
No. It explains under what conditions the mechanism could be maintained by selection. The connection of this mechanism with experience requires a separate justification.

2. Why don't closely related species provide completely independent evidence?
Their common features can be traced back to a single historical event. The comparative analysis must take into account the overall history, otherwise the amount of independent information will be overstated.

3. Why save multiple origin stories?
Given incomplete data, they may equally explain the observed distribution of traits. The apparent alternatives show which additional observations will truly differentiate the reconstructions and which conclusions are still premature.

Return to chapter 19

Chapter 20 Analysis of practice

You cannot demand that a small project immediately solve all questions of consciousness. But its output must correspond exactly to the work performed. A strong principle takes on more obligations than the prediction of a report; it must be justified consistently, without arbitrarily excluding inconvenient implementations.

Self-test answers

1. Why might an explanation of color discrimination fail to address the question of experience?
It shows how the system classifies and matches signals. The claim that these processes are experiential requires a separate connection between computational organization and phenomenal property.

2. Is it necessary to introduce an additional entity to explain the experience?
No. A strategy of physical identity is possible. But she must justify which process is identical to experience and why, and not simply declare the two descriptions to be identical.

3. What is considered progress if there is no definitive explanation yet?
A new discriminating prediction, the elimination of a serious alternative, a more precise correspondence between the structure of the mechanism and the structure of experience, and a clear narrowing of the remaining assumptions. These results can be verified independently of the claim to solve the whole question.

Return to chapter 20

Chapter 21 Analysis of practice

To study memory, we can conclude that there is a certain dependence of choice on available information under given conditions. For damage research, exposure registration, avoidance, and perceived negative experiences remain at different levels. The closest alternative must be specified by a rule or mechanism, and not by the word “instinct.”

Self-test answers

1. Why does recognizing possible pain not mean recognizing explicit reflection?
Negative experience, control of the state of the body and consideration of one’s own reasons are different properties. Data about the first does not automatically become evidence of the others.

2. Is it possible to extend the result of one species to all its relatives?
Only with additional justification, taking into account the mechanisms, relationships and conditions of the task. The proximity of species helps to build a hypothesis, but does not replace testing its boundaries.

3. What to do with an explanation through a simple rule?
You need to write it down accurately and check it against the same data and new conditions. If the rule explains the outcome, the more complex inference is weakened; if it is systematically wrong, there are grounds for a richer model.

Return to chapter 21

Chapter 22 Analysis of practice

Until the new B, the reactions are the same. On B they are equal to 1 and 0.32768, so observation separates the rules. After five pause steps, the answer to A in a specific model is about 0.779695, the resource has about 0.758219. These differences are consequences of a given calculation and do not establish qualia for the system.

```
Reproduction: python teaching_lab.py --chapter 22
```

Self-test answers

1. Can functional memory exist without a nervous system?
Yes, the state of a system can retain the influence of past influences and change future responses. This is a broad functional sense of memory, which should be distinguished from experienced memory.

2. Why doesn't a decrease in response prove learning?
It can be caused by exhaustion, corruption, changes in input availability, or other processes. Control conditions, tests of recovery, specificity, and a more accurate model of alternatives are needed.

3. Does the absence of neurons make the question of qualia meaningless?
No. It changes the range of possible mechanisms and the requirements for evidence. A theory that allows for a different medium must specify a sufficient organization and method of verification, and not be limited by the overall complexity of the system.

Return to chapter 22

Chapter 23 Analysis of practice

Changing an independent template can change the text without changing the calculations. A target change in the used state can change actions, but its role must be checked against the rules of a particular implementation. The diagram supports questions about architecture and access; in itself it does not justify the experienced confusion or other quality.

Self-test answers

1. Why is successful conversation about experiences not sufficient to draw conclusions about them?
Text can be generated by a mechanism unrelated to the internal states it describes. We must investigate how the response arises, what information is accessible, and what causal role the hypothesized self-control plays.

2. Does a biological bearer guarantee consciousness?

No. The material does not automatically determine the organization required by the theory under consideration. It is necessary to study a specific system, its connections and interaction with external control.

3. Why is a candidate’s profile better than the number of characteristics completed? It preserves the dependence of the conclusion on the theory and distinguishes between necessary, sufficient and concomitant. A simple amount can hide the absence of a critical condition and create unreasonable precision in the estimate.

Return to chapter 23

Chapter 24 Analysis of practice

A separate history may be useful, but a match with a common tracker will limit the output to the identity mechanism. The lack of discriminative conditions makes even a large sample uninformative regarding this issue. The finished specification should show what data and what computable rules are needed next.

Self-test answers

1. History can be used by a general mechanism for learning and tracking objects. For a special output you need a corresponding competitor.
2. This comparison confuses the structure of the model with the advantage of available information; the contribution of the mechanism itself has not yet been identified.
3. Otherwise, the hypothesis begins to serve as its own proof. The division shows what the model must already take into account and what remains to be verified.
Return to chapter 24

Chapter 25 Analysis of practice

M0 gives 0.75; 0.25; 0.75; 0.75. M1 gives 0.75; 0.4375; approximately 0.710526; approximately 0.858108. M2 and the general tracker give 0.75; 0.25; 0.875; 0.5625. When holding zero, all updates start from 0.5. Simultaneously renaming sources changes the designations of the records, but not the numerical relationships.

```
Reproduction: python teaching_lab.py --chapter 25
```

Self-test answers

1. It returns to the history of A, keeping its confirmation separate from B's contradiction.
2. In this implementation, M1 and M2 will coincide, so the value of individual stories cannot be verified.
3. The answer also depends on the scale and how it is used. The connection between the calculation and the report of the experience cannot be considered as given by the name of the variable itself.

Return to chapter 25

Chapter 26 Analysis of practice

Independent recording may reproduce the usual trajectory and may not preserve targeted responses to interventions. The one-way controller requires a separate missing direction test. A simple regulator cannot be ruled out by the mere impression of simplicity: the substantive conditions of the definition must be checked. None of these tests by itself proves that the organization is sufficient for the experience.

```
Reproduction: python teaching_lab.py --chapter 26
```

Self-test answers

1. It asserts the sufficiency of a particular organization for the existence of experience; predictive reports establish a more limited relationship.
2. What is important in the hypothesis are the relationships and transitions preserved by such renaming.
3. No. The assembly checks the implementation of the conditions. Sufficiency, boundaries of implementation and transfer to a phenomenal conclusion require their own justification.

Return to chapter 26

Chapter 27 Analysis of practice

M0 uses the current window, M1 the general history, M2 the history of the recognized source. If the general tracker matches M2, winning over M1 supports individual histories but does not highlight a special bodily organization. The phenomenal implications of the principle require independent verification of qualitative relationships and do not follow from functional comparison alone.

Self-test answers

1. No. Need different models or other comparison item.
2. It shows the role of current causation relative to weak control, but does not exclude full-fledged alternative mechanisms.
3. Describe it as discovered in the analysis, derive a new prediction, and test it on independent data.

Return to chapter 27

Chapter 28 Analysis of practice

The basic full mode has a success rate of about 0.8794 and a utility of 0.8619 per step. The fixed model has an average success of about 0.4997, hiding 0.8780 before and 0.1213 after the inversion. The difference supports the functional

benefit of the management update in a given environment. Utility and accuracy vary due to observational costs.

Self-test answers

1. We disable one specific mechanism. Removing all memory at once would change several reasons for the outcome.
2. The average combines almost correct control and almost systematic error after switching.
3. The forecast would cease to measure uncertainty before the outcome. There would be a leakage of the answer into the metric being assessed.
Return to chapter 28

Chapter 29 Analysis of practice

Rotation and translation preserve distances with numerical precision; stretching usually changes them. Recurrent steps and the same unfolded layers coincide. In the delivery problem, full selection and direct access give 100 percent: this shows the work of the channel, and not its exceptional need or phenomenality.

Self-test answers

1. The selected transformation preserves the geometry of the given model. The presence of experience does not follow from this.
2. Specific final calculations have been verified. A theory may make additional commitments about time, state, or physical implementation that require separate testing.
3. For this task, the required information can be used in at least two ways. The contribution of the limited general space has not yet been identified.
Return to chapter 29

Chapter 30 Analysis of practice

M2 matches the general tracker according to the given rules. Direct internal sensor has access by design. An adaptive strategy changes actions based on past mistakes. The accelerated revision of the base composite candidate has zero inclusions. These are different outcomes requiring different next steps; they cannot be combined to prove the experience of one system.

Self-test answers

1. It shows that minimal calculations do not yet distinguish between special self-reference and general accounting of source history.
2. Operation of a given access channel and measurement verification. Stronger conclusions require additional support.

3. No. The mechanism was not activated in the baseline scenario; it needs a predetermined experiment that will test its predicted role.
Return to chapter 30

Chapter 31 Analysis of practice

The training table retains the original IDs and explicit exceptions. The actual structure of the downloaded file is not replaced by a brief description of the repository. Added STTC can improve or worsen the error; valid output is specific to the selected procedure and this set. Testing by batches is stricter than testing by organoids regarding batch transfer, but also does not eliminate all dependencies and does not establish experience.

```
Reproduction: python teaching_lab.py --chapter 31
```

Self-test answers

1. The result can be jointly provided by the tissue, the input transformation, the feedback algorithm, and the readout.
2. The first evaluates the temporal transfer within a sample, the second evaluates the transfer between samples.
3. No. It establishes a certain functional result; the phenomenal inference requires an additional argument about the mechanism and the correspondence principle.
Return to chapter 31

Chapter 32 Analysis of practice

The profile must distinguish between implemented and verified, implemented without a verified role, and unknown. The absence of a shutdown effect requires checking activation, task sensitivity, and compensation. Even a complete functional profile does not independently confirm the sufficiency of its own theoretical requirements.

Self-test answers

1. No. It is necessary to check activation, information dependencies and impact on the decision.
2. The first shows the possibilities of external analysis, the second shows the use of information by the system itself.
3. The implementation tests the candidate design, and sufficiency is a separate theoretical hypothesis and needs independent reasons.
Return to chapter 32

Chapter 33 Analysis of practice

Recovery does not have to be error-free. The result includes the generation rule, comparison criterion, all selection outcomes and uncertainty. The same type of additional events does not eliminate structural non-identification. Winning among bad models can be combined with a poor absolute forecast.

```
Reproduction: python teaching_lab.py --chapter 33
```

In the attached run, 30 sequences of 120 events are built for each generator. The selection is made based on the smallest Brier error of the observation forecast; the candidate parameters are fixed here. The row of the matrix shows the given generator, the column shows the selected candidate. This is a diagnostic teaching experiment, not a comparison based on unknown human nature.

Generator	Selected M0	Selected M1	Selected M2
M0	25	1	4
M1	3	24	3
M2	3	5	22

On the selected grid, the smallest Brier error 0.231314 was given by a pair of parameters: history preservation 0.80, observation reliability 0.75. In this run, it coincided with the pair assigned to the generator. One such result is not an estimate of the accuracy of restoration in all conditions.

Self-test answers

1. The sign is noisy and can occur for various hidden reasons; We need a forecast for both possibilities.
2. No. It tests the analysis in a known artificial situation.
3. The rule specifies causal dynamics and new predictions. Subtle substitution makes it difficult to assess what was actually tested.

Return to chapter 33

Chapter 34 Analysis of practice

The protocol must contain specific data, the order of calculations, the main result and conditions of the solution. The reference point 0.01 is a selected educational criterion, and not a universal scientific norm. Stopping after the first successful distinction without a predetermined sequential procedure changes the meaning of the test.

```
Reproduction: python teaching_lab.py --chapter 34
```

For 100 independent sequences with two sources, the average pairwise error difference M1 minus M2 is equal to 0.016965; percentile interval 95 percent with 4000 repeated samples of whole pairs: [0.013687; 0.020376]. The pre-selected

training threshold is 0.01. The interval refers to the generator and procedure of this experiment. The protocol is saved until the result is calculated; This is a local record and not a public record of independent research.

Self-test answers

1. No. This analysis must be distinguished from a pre-planned inspection.
2. Otherwise, the answer being tested itself may be included in the forecast.
3. No. This is the selected scope of the case study; for a specific study, the volume is justified by the required accuracy and the difference in models.
Return to chapter 34

Chapter 35 Analysis of practice

The average 0.008 in the educational illustration does not in itself answer the question about the practically interesting advantage of 0.01. Version matching with zero inclusions of a component does not check its operation in active mode. The choice of a strong next experiment is the meaningful result of the analysis.

Self-test answers

1. The completed task did not check the consequences of running this module; to evaluate its contribution, discriminating conditions are needed.
2. The same result can be given by the general memory of the identity of the sources M1I, which coincides with M2 according to the rules.
3. The revision specifies an explicit new rule and consequences for the independent conditions; adjustment may be limited to improving the result already seen.
Return to chapter 35

Chapter 36 Analysis of practice

A full-fledged work allows you to repeat the path from inputs to output and does not require agreement with the phenomenal hypothesis to verify the calculations. Independent contribution begins with a new discernible consequence. If the proposed principle of sufficiency is not yet justified, this is clearly indicated and linked to a specific next study.

Self-test answers

1. He established the consequences of the chosen separate memory mechanism and its coincidence with the general object tracker.
2. The working hypothesis requires mutual causal influence; general activity or signal transmission alone does not establish it.
3. Yes. It must clearly define the test performed, preserve the research objective, and suggest the next discriminating step.
Return to chapter 36

Sources

[1] Dennett, Daniel C. (1988). Quining Qualia. In A. J. Marcel and E. Bisiach (eds.), Consciousness in Contemporary Science. Oxford University Press
[2] Chalmers, David J. (1995). Facing Up to the Problem of Consciousness. Journal of Consciousness Studies, 2(3), 200–219
[3] Shepard, Roger N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. Psychometrika, 27, 125–140
[4] Kalckert, Andreas; Ehrsson, H. Henrik (2012). Moving a Rubber Hand that Feels Like Your Own: A Dissociation of Ownership and Agency. Frontiers in Human Neuroscience, 6, 40
[5] Schwitzgebel, Eric (2008). The Unreliability of Naive Introspection. The Philosophical Review, 117(2), 245–273
[6] Fleming, Stephen M.; Lau, Hakwan C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8, 443
[7] Machamer, Peter; Darden, Lindley; Craver, Carl F. (2000). Thinking about Mechanisms. Philosophy of Science, 67(1), 1–25
[8] Petitmengin, Claire (2006). Describing one’s subjective experience in the second person: An interview method for the science of consciousness. Phenomenology and the Cognitive Sciences, 5, 229–269
[9] Logothetis, Nikos K.; Pauls, Jon; Augath, Mark; Trinath, Torsten; Oeltermann, Axel (2001). Neurophysiological investigation of the basis of the fMRI signal. Nature, 412, 150–157
[10] Nosek, Brian A.; Ebersole, Charles R.; DeHaven, Alexander C.; Mellor, David T. (2018). The preregistration revolution. Proceedings of the National Academy of Sciences, 115(11), 2600–2606
[11] Simmons, Joseph P.; Nelson, Leif D.; Simonsohn, Uri (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11), 1359–1366
[12] Hodgkin, Alan L.; Huxley, Andrew F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. The Journal of Physiology, 117(4), 500–544
[13] Sperling, George (1960). The information available in brief visual presentations. Psychological Monographs: General and Applied, 74(11), 1–29
[14] Simons, Daniel J.; Chabris, Christopher F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception, 28(9), 1059–1074
[15] Samad, Majed; Chung, Albert Jin; Shams, Ladan (2015). Perception of Body Ownership Is Driven by Bayesian Sensory Inference. PLOS ONE, 10(2), e0117178
[16] Dehaene, S.; Kerszberg, M.; Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. Proceedings of the National Academy of Sciences 95(24):14529–14534
[17] Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. Trends in Cognitive Sciences 10(11):494–501
[18] Fahrenfort, J. J.; Scholte, H. S.; Lamme, V. A. F. (2007). Masking disrupts reentrant processing in human visual cortex. Journal of Cognitive Neuroscience 19(9):1488–1497
[19] Albantakis, L.; Barbosa, L.; Findlay, G.; Grasso, M.; Haun, A. M.; Marshall, W.; Mayner, W. G. P.; Zaeemzadeh, A.; Boly, M.; Juel, B. E.; Sasai, S.; Fujii, K.; David, I.; Hendren, J.; Lang, J. P.; Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. PLOS Computational Biology 19(10):e1011465
[20] Lau, H. C. (2008). A higher order Bayesian decision theory of consciousness. Progress in Brain Research 168:35–48

[21] Rao, R. P. N.; Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience 2(1):79–87
[22] Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. Trends in Cognitive Sciences 17(11):565–573
[23] Cogitate Consortium; Ferrante, O.; Górska-Klimowska, U.; Henin, S.; Hirschhorn, R.; et al. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. Nature 642:133–142
[24] Cohen, M. A.; Dembski, C.; Ortego, K.; Steinhibler, C.; Pitts, M. (2024). Neural signatures of visual awareness independent of postperceptual processing. Cerebral Cortex 34(11):bhae415
[25] He, L.; Zhang, Z.; Wang, Y.; Meng, Q.; Jiang, Y.; Xu, W.; Wu, J.; Sheng, J.; Ye, Z.; Ding, X. (2026). Report-modulated prefrontal activity and consistent posterior representations during conscious visual perception. Consciousness and Cognition 142:104075
[26] Zhang, Y.; Li, X.; Jin, Z.; Zhang, J.; Li, L. (2026). Distinct frontal lobe subregions mediate the emergence and reporting of visual consciousness. NeuroImage 334:121964
[27] Noorman, S.; Stein, T.; Zantvoord, J.; Fahrenfort, J.; van Gaal, S. (2025). A causal role of the NMDA receptor in recurrent processing during perceptual integration. eLife 13:RP100530
[28] Kouider, S.; Stahlhut, C.; Gelskov, S. V.; Barbosa, L. S.; Dutat, M.; de Gardelle, V.; Christophe, A.; Dehaene, S.; Dehaene-Lambertz, G. (2013). A neural marker of perceptual consciousness in infants. Science 340(6130):376–380
[29] Amsterdam, B. (1972). Mirror self-image reactions before age two. Developmental Psychobiology 5(4):297–305
[30] Goupil, L.; Romand-Monnier, M.; Kouider, S. (2016). Infants ask for help when they know they don't know. Proceedings of the National Academy of Sciences 113(13):3492–3496
[31] Felsenstein, J. (1985). Phylogenies and the Comparative Method. The American Naturalist 125(1):1–15
[32] Barron, A. B.; Klein, C. (2016). What insects can tell us about the origins of consciousness. Proceedings of the National Academy of Sciences 113(18):4900–4908
[33] Andrews, K.; Birch, J.; Sebo, J.; et al.; individual signs (2024). The New York Declaration on Animal Consciousness. Official statement, 19 April 2024
[34] Hampton, R. R. (2001). Rhesus monkeys know when they remember. Proceedings of the National Academy of Sciences 98(9):5359–5362
[35] Nieder, A.; Wagener, L.; Rinnert, P. (2020). A neural correlate of sensory consciousness in a corvid bird. Science 369(6511):1626–1629
[36] Crook, R. J. (2021). Behavioral and neurophysiological evidence suggests affective pain experience in octopus. iScience 24(3):102229
[37] Barr, S.; Elwood, R. W. (2024). Trade-Offs between Avoidance of Noxious Electric Shock and Avoidance of Bright Light in Shore Crabs Are Consistent with Predictions of Pain. Animals 14(5):770
[38] Manzi, O.; Lynch, K. E.; Allman, D. M.; Latty, T.; White, T. E. (2026). Flexible self-protection as evidence of pain-like states in house crickets. Proceedings of the Royal Society B 293(2070):20260609
[39] Toyota, M.; Spencer, D.; Sawai-Toyota, S.; et al. (2018). Glutamate triggers long-distance, calcium-based plant defense signaling. Science 361(6407):1112–1115
[40] Boisseau, R. P.; Vogel, D.; Dussutour, A. (2016). Habituation in non-neural organisms: evidence from slime moulds. Proceedings of the Royal Society B 283(1829):20160446
[41] Markel, K. (2020). Lack of evidence for associative learning in pea plants. eLife 9:e57614
[42] Butlin, P.; Long, R.; Bayne, T.; Bengio, Y.; Birch, J.; Chalmers, D.; Constant, A.; Deane, G.; Elmoznino, E.; Fleming, S. M.; Ji, X.; Kanai, R.; Klein, C.; Lindsay, G.; Michel, M.; Mudrik, L.; Peters, M. A. K.; Schwitzgebel, E.; Simon, J.; VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. Trends in Cognitive Sciences 30(6):488–501

[43] Wilson R. C., Collins A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. eLife 8, e49547
[44] Kagan B. J. et al. (2022). In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. Neuron 110(23), 3952–3969.e8
[45] Cai H. et al. (2023). Brain organoid reservoir computing for artificial intelligence. Nature Electronics 6, 1032–1039
[46] Robbins A. et al. (2026). Goal-directed learning in cortical organoids. Cell Reports 45(2), 116984
[47] Robbins A. (2025). Goal-Directed Learning in Cortical Organoids — Experimental Data and Code. Zenodo, version v2, doi:10.5281/zenodo.17684862
[48] International Society for Stem Cell Research (2025). Guidelines for Stem Cell Research and Clinical Translation. Version 1.2, August 2025 update
[49] Lindsey J. (2025). Emergent Introspective Awareness in Large Language Models. Transformer Circuits Research Report, 29 October 2025; preprint arXiv:2601.01828, 5 January 2026

Author and projects

Eudgene Brenev

To report an error found, discuss the material, or get acquainted with the author's projects, use these addresses.

EMAIL

info@brenev.com

AUTHOR'S WEBSITE

brenev.com

MIRA WEBSITE

mira.technology

OTHER PROJECTS

haraizen.tech

TELEGRAM

https://t.me/Eudgenee

The title of the book refers to the author’s artificial intelligence project “Mira”.