About this course

• an opinionated survey of topics in philosophy, psychology, and neuroscience (plus a bit of computer science and linguistics)

• by the end of the course, you'll have one big picture of the

nature of the mind, and some evidence for and against that big

picture

Three Questions

1. How should we study the mind?
2. What is consciousness?
3. What is rational action?

How should we study the mind?
Answer 1: We should study the mind as a system of representations and of computations over those representations.

what does it even mean to say that the mind is a system of representations and computations over those representations?
our class theory says that the mind is, fundamentally, a mechanism like clockwork or hydraulics
the "pieces" are representations, and that the way they "move" are the computations performed over those representations
the more modern analogy is to computing devices the mind's operations are like those of a clockwork or hydraulic mechanism, but they are even more like those of an electronic computer
In our approach, saying that the mind is like a computer is not a metaphor mental processes are literally computational processes, taking representations as inputs and producing representations as outputs how could the mind literally be a computational system? and why would anyone think that it is?•

1. a lot of mental states seem to be representations
2. a lot of mental processes seem to be computations

first idea: a lot of mental states seem to be representations
ordinary representations are things like maps, speedometers, representational artworks,blueprints, traffic lights, and so on.

key features of representations:

1. they are about things, the way maps and blueprints are about things.
2. what they are about is some conceivable state of affairs; in philosophical jargon, this is their "content" that content can even beimpossible

3. some representations are true or false, while others are obeyed or disobeyed

according to linguists, ordinary sentences are representations:

"Tim put the kitten next to the food."
indicative representation
-is about Tim, the kitten, its placement, and the food
-content is that Tim placed the kitten next to the food some time in the past
-can be true or false

"Tim, put the kitten next to the food."
-imperative representation
-is about Tim, the kitten, its placement, and the food
-content is that Tim places the kitten next to the food sometime in the future
- tells Tim to do what's in the content
-can be obeyed or disobeyed

Three key features of representations;
-Representations can be indicative or imperative
-indicative representations say how things are
    e.g., maps, descriptions
-imperative representations command that things be a certain way
    e.g., blueprints, instructions

Representations have contents
-what an indicative representation says is true is its content
-what an imperative representation commands to be made true is its content

The contents of some representations are impossible
-it's impossible for Santa to visit my house, because there is no such person as Santa
    yet I can create a representation that he visited my house
-it's impossible to make a round bowl with a circumference exactly three times the diameter, because pi is not exactly 3
    yet I can command someone to make me such a bowl

So why think that mental states are also representations?

consider some mental states:
-I believe that I live in Houston;
- I remember that my father likes buttersoup
- I am afraid that one day my wife and son will get hurt in a collision..

my belief that I live in Houston has all the key features of a representation
-it is about Houston, me, residency
-it has a content: that I live in Houston
-the content can be true or false
-the content is possible (and other of my beliefs, like my childhood beliefs about Santa and infinity, were impossible)

-my belief is like an indicative representation in that it can be true or false

-similar things hold for my memory, my fear...
-most things in the world do not have the features of representations
    -most rock formations are not “about” anything, do not say anything at all, are not true or false,obeyed or disobeyed
    -coral reefs are not “about” water conditions or marine life, cannot lie about past ocean conditions
    -clouds can look like, e.g., rabbits, but cannot misrepresent what rabbits really look like and get them wrong
    -almost nothing in the natural world is like a representation, except for mental states
    -so maybe mental states are a kind of representation, too

second idea: a lot of mental processes seem to be computations
familiar computations include things like computing how many eggs to putinto a recipe that you’re halving, computing website ranking functions, and computing how you could theoretically complete your Rice degree in just another three semesters plus a summer course we sometimes call these ‘calculations’ instead of ‘computations’, but it seems like we’re talking about the same thing using either term

key features of computations:
-computations are processes that take time
-they have inputs and transform them into outputs•
-the inputs and outputs are both representations
-there are patterns to these transformations; not just any transformation would count
-the pattern of the transformation can be characterized in mathematical or logical terms

so why think that mental processes are also computational processes?

•consider some mental processes:

extracting information from a visual scene, remembering a name,choosing an option that will get you what you really want...

these mental processes have the features of computational processes

when I'm extracting information from a visual scene, the input to the process is the stimulation of my eyes by light from the scene, and the output is a belief that yes, there's a dog present (or whatever information I'm extracting)

- the process takes time, of course, and is non- random
- it operates on representations of incident light (and maybe other intermediate representations) and generates a new representation (in belief)
- if the last seventy years of theoretical research in vision have been on the right path, there is also some very complex mathematical or logical way of describing how to get from those sorts of inputs to those sorts of outputs

it is perhaps not so amazing that mental processes take time, or have stable relations between how they start and how they end most processes are like this, and most processes can be given mathematical or logical descriptions. But most processes do not take representations as inputs or produce representations as outputs. One more feature of mental processes is striking: so many of our mental processes are logical or mathematical in their own right they involve logical inference or solving mathematical problems (whether consciously or unconsciously). They are processes that seem to involve USING logic and math in order to execute them that's quite different from most natural processes so: many mental states seem to be representations, and many mental processes seem to be computations. The mind seems to be like a digital computer this is an exciting starting point, but it faces many challenges!

We'll look at those challenges next

Materialism:

materialism is the theory that there is just one fundamental kind of stuff ("matter"), and every feature of the world is some arrangement of that stuff: gold, hurricanes, live oak trees, human happiness, etc

Our approach to studying the mind holds that the mind can be treated as a mechanical system, like a clockwork or computer. Representations, made of neurons, cause the formation of other representations.

This pretty much presupposes materialism. We should do a little to examine our presupposition

Scientific challenges to materialism

What would you say is the hardest thing about the mind to explain in materialistic terms?

Many theorists have been skeptical that there are full scientific explanations for:

creativity, love, rational choice,irrational choice, moral sensibilities, consciousness...

Are there any specifically scientific arguments that there are limits to materialistic explanations of the mind? These would be arguments based on incompatibilities between the scientifically investigated features of matter and the scientifically investigated features of the human mind

Let’s consider one possible argument, concerning creativity.

- creativity is something like: coming up with new ideas that are not trivially derivedfrom ideas you got elsewhere
- can be in any domain: ideas for recipes, dances, experiments, sculptures, arguments, political systems, currencies… the challenge
- matter is inherently bound by the laws of physics, while human creativity seemingly reveals capabilities for breaking rules, doing the unexpected, and violating previous patterns.

However, giving a materialist theory of creativity might seem easy

- don’t we have AI models of creativity already?
- what exists, and what doesn’t exist, right now?

  a few AI models of creativity:
    - creative playing of chess, go, computer games…
    - creating new essays and digital art from prompts...
    - guessing 3D protein folding from just the amino acid sequence etc.

Materialism

- people say, “computers only dowhat you tell them to”
- that’s supposed to show that AI creativity isn’t “real” creativity if you make AI systems that can learn, then it seems hard to take seriously the objection that an AI only does what you tell it to do
- it only learns the way you tell it to learn, but people only learn the way our brains allow us to learn, so that would not make us much different what do you think?
- is there still an aspect of human creativity that escapes scientific explanation?
- is there an aspect of human creativity that is inconsistent with what we know about the physical, chemical, biological world in general?

In our third module, we will at least try to understand rational and irrational action, which might make a good start toward understanding creativity.

ORGANIZING THE MIND

our hypothesis says that the mind is best studied as a system of representations and computations over those representations. But, “the mind” covers a lot of stuff. We should consider whether all of the mind counts as representations or computations over representations

THE MIND

Anger,Planning,Wishing,Wanting,Expectation,Dream,Hope,Seeing,Hearing,Imagining,Smelling,Tasting,Bias,Re
dge,Thought,Feeling,Cruelty,Hunger,Thirst,Joy,Sorrow,Anger,Planning,Wishing,Wanting,Expectation,Dream,Ho
prioception,Uncertainty,Contentment,Jealousy

That list was far from complete, but I hope you’ll agree that it was at least pretty representative. It didn’t leave out major categories of mental states or mental processes. Does the list reveal that the mind is, as claimed, made up of representations and computations?

The main parts of the mind are often held to be:
Perceptions,beliefs,memories, - Indicative representations
Desires,plans,choices - Imperative representations

perceptions, beliefs, and memories are all indicative representations because

- they are about things
- they are sometimes about things that are not real, or even impossible
- they are all true/accurate or false/inaccurate

Desires, plans, and choices are imperative representations because:

- they are about things
- they are sometimes about things that are not real, even impossible
- they can be obeyed/executed successfully or disobeyed executed unsuccessfully

The Mind: perception
Proprioception, pain, pleasure, tasting, smelling, imagining, hearing, seeing, feeling

The Mind: belief and memory
Knowledge, expectation, bias, impression, memory, uncertainty

The Mind: desire
Hunger, thirst, wishing, wanting, desire

The Mind: planning and choice
Planning, intention

There are also cognitive processes, that is, processes taking us from sensory stimulation to perception to current belief to memory (and maybe more) and conative processes, that is, processes taking us from desire to plans to choices to movements (and maybe more)

The Mind: cognitive and conative processes
Thought, reasoning, inference, association, language, imagining, planning

Here are also mental states and processes that involve both indicative and imperative representations Some think emotions are in this category. Some think that character traits are in this category

The Mind: emotions and traits
Joy, sorrow, anger, hope, kindness, angst, horror, delight, contentment, jealousy

This covers almost everything brought up so far

- does our system have a place for each mental state and process to fit into?

- what might be missing from our discussion?

Representation

If the mind is a system of representations and computations over those representations, then we need a theory of representation.
-It must be a good theory of familiar representations.
-It must not build representations out of any psychological states or processes
-It must make it possible that brains contain representations

reminder: representations have certain features

- they have intentionality: they are about things
- they are indicative (true/false) or imperative (obeyed/ disobeyed)
- they have a content: a state of affairs that they say is true or demand be made the case that a materialist can't say that mental states are made from representations, while also saying that representations are made from mental states such as purposes.

As philosopher Fred Dretske writes, you can't make cake from cake. Meaning, you can't show that a mind is made of matter by showing that it is made of X, Y, Z, when Z is something mental (like a purpose). Any theory of representation we consider has to pass this test:

- it cannot presuppose human (or other) purposes, or any other mental states

So, we need a theory of representation that is correct, but that does not bring up human purposes some human-made representations are pictures. Can we just run with this observation?

- For X to represent Y is for X to be a picture of Y

Does this pictorial theory of representation try to make cake from cake?
probably not

- At least some pictures have a clear objective relation to reality (e.g., they reflect light in a pattern similar to thepattern of light reflected by a real thing)that exists whether or not anyone intends for that relation to exist but there aren't literal pictures in the brain

So this isn't a theory of representation we can use. If a pictorial theory is the right theory of representation, then our big picture is just wrong! However, there are lots of ordinary representations that are not like pictures. e.g., sonic representations that don't sound like what they represent

- If there are neural representations, they are going to be like fire alarms: not at all pictorial

A more sophisticated approach might be to develop a theory based on isomorphism

- an abstract sort of mapping between one thing and another after all, there's an obvious mapping between a picture and what it's a picture of and a less-obvious, but still real, mapping between an alarm's ringing and there being a fire. For mathematicians (and philosophers), an isomorphism is a very specific sort of relationship between two collections

For one set of objects, having one set of properties (and standing in one set of relations) to be isomorphic to another, it must be possible to make a mapping between the objects in the first set and the objects in the second set, such that:

1. The mapping maps each object in the first set onto a distinct object in the second set and leaves no unmapped objects in the second set.
2. If the objects in the first set include some with a property, F, or stand in some relation, R, to one another, then their mapped-onto counterparts in the second set will also have a parallel property, G, or stand in some parallel relation, S, to one another. The sense of “parallel” here is that, if, say, objects a, b, and c in the first set have property F, or stand in relation R to one another, then their mapped-onto objects all have some property G, or stand in relation S to one another. Likewise, if a, b, and c don’t stand in R, then neither do their mapped-onto objects, and so on.

This probably sounds pretty confusing, but its application is not so very hard to understand in the case of the mind. The idea would be that neural states represent states of the world by being isomorphic to them. For instance, there might be a group of neurons in the visual cortex (that’s the first set of things) that fire (that’s the property they have) in response to the presence of light/dark boundaries (that’s the mapped-onto property) in specific regions of space in front of the viewer (those regions are the second set of things). There are also well-known mappings between activity in the auditory cortex and volume/pitch, and between activity in the somatosensory cortex and touch/pressure/location. It’s a little less clear what the mappings might be between neural structures and, e.g., all the things you know about all the people you know. Is there a neuron in you that is mapped onto Beyoncé? Maybe there is! Maybe there is a neuron (or neural structure) that is mapped onto Beyoncé, and another that is mapped onto being from a place, and another that is mapped onto Houston. Activating them together is then mapped onto the fact that Beyoncé is from Houston. Ludwig Wittgenstein famously defended something like this view of representation at the beginning of the 20th century; other philosophers have worked on it more recently. Unfortunately, there are some huge problems for the isomorphism theory of representation.

Some problems:

- Isomorphism is symmetrical; representation is not.
- There are just too many isomorphisms out there.
- You can’t have an isomorphism between a set of neurons and a set of things that don’t exist.
- So, the isomorphism theory of representation is wrong.
- We need a new theory.

Representation: Teleosemantics

Today we'll look at one theory that avoids the problems of pictorial and isomorphism theories of representation.

- There is a family of theories called 'teleosemantic' theories of representation, developed by philosophers including David Papineau, Karen Neander, and Fred Dretske.
- My favorite version comes from Ruth Millikan.
- Millikan's theory of representation is called a teleosemantic theory because her theory of how a meaningful (semantic) entity, a representation, comes into existence relies on an appeal to purpose or function (telos).

Millikan's Theory of Representation:

- An indicative representation is a structure that must map onto the world in a certain way in order for some other structure to perform its function.
- An imperative representation is a structure such that some other structure performs its function by changing the world so that the first structures will then map (in a certain way) onto the new state of the world.
- An indicative representation must actually map onto the world in a certain way for some other structure to perform its function.
- An imperative representation allows some other structure to perform its function by generating a mapping between the first structure and the world in a certain way.
- Both parts of the theory rely on the idea that structures in the brain perform functions.
    - By saying that representations help us carry out functions, Millikan makes it clear how a false representation is an error or mistake: it's a representation not contributing to the job it has to do.
    - BUT! These functions had better not come from our own purposes, or else we are trying to make cake from cake.

Representation: Function

- Millikan has a view of function where having a function does not require being given a function by a person.
- Instead, on her view, functions are derived from histories of natural selection.
    - Consider the heart:
        - The function of the heart is to pump blood.
        - Why? The function of the heart is to pump blood because pumping blood is what past hearts did that caused the genes producing them to be selected over alternative genes.
        - The heart is "designed" by nature for pumping blood, and that metaphorical design gives the heart a literal function.

- Now consider a neuron in area V1 of the visual cortex:
    - Its activity might or might not map onto the presence of (say) a light/dark boundary at a certain location in space.
    - If it does, then some other neuron can do its job. But what other job?
        - Actually, there are many other jobs to do, such as: generating mappings onto movement (MT), mappings onto color (V4), preparing the body for eye movements (FEF), etc.
        - There is a history of natural selection building brains in one way rather than another, out of alternatives, because brains built in the one way had features that were selected for.
        - Thus, the brain has many functions to perform, functions given to it by its history of natural selection.

- Therefore, neural firing A mapping onto state of the world B can be necessary in order for some structure in the brain to perform its function.
- Thus, neural firing A can represent that the world is in state B.

Computation

Our class theory suggests that the mind is a system of representations and computations over those representations. We've thought a lot about representations, but what about computations?

Objections to Mental Computation

Many people reject the idea that mental processes are computations. Why might they do so?
-It doesn’t seem consistent with experience.
-It doesn’t seem like a clear hypothesis.
Today, we’ll discuss the first group of objections: those that argue mental processes are not computations because they don’t seem consistent with experience.

---

1. The Experience of Computing:

In general, it’s very rare to feel like your mind is "computing." There are, however, some special cases.

-For example, when you work through a complicated arithmetic problem like 234 - 76 = ?, it likely feels like you're computing the answer. In this case, your mental process feels computational.

-But what's striking is that most of our mental processes don’t feel like this at all. For instance:

Seeing how far away a car is.

Remembering the name of an acquaintance.

Understanding a sentence like "Visiting relatives can be boring."

These processes feel radically different from doing mental arithmetic.

---

2. The Speed of Computation:

-Can we just say that most computational processes happen too quickly to notice?

-This seems plausible in some cases. For instance, visual researchers suggest that visual representations of edges are used to compute representations of shape, and so on. These computational processes occur in just a few milliseconds.

-Visual stimuli that last less than 25-50 milliseconds don’t make it into consciousness, according to psychologists. So, it’s possible that computational processes that happen in this timeframe also don’t make it into conscious awareness.

---

3. The Tip-of-the-Tongue Phenomenon:

-But not all mental processes are that fast. For example, if you have a word on the “tip of your tongue” and then suddenly remember it, the process of retrieval might take a second or so.

-If speed is the only factor hiding the computational nature of our mental processes, we should notice that retrieving a word feels like computing. But it doesn’t feel like that at all.

---

4. The Role of Voluntary Action:

-Here’s a different thought: perhaps our voluntary mental processes don’t feel like computations because they’re voluntary. When I focus my attention, reason through a problem, plan for the future, or try to suppress (or encourage) an emotion, I feel like I am the agent of these mental processes.

-When my own mental actions are part of my mental processes, it doesn’t feel like there is any "computation" going on that I’m merely executing.

-Free will plays a role here. Our sense of free will partly comes from the feeling of not following a set program. While there’s plenty of debate about whether we genuinely have free will, there’s very little debate about the fact that we feel free.

5. The Computational Theorist's Reply:

How should a computational theorist of the mind reply to these objections?

First, computational processes in your own mind are not generally noticed by you because they are the processes through which you notice (or otherwise represent) anything at all.

Second, not noticing computational processes, or feeling sure they're absent, isn't evidence for their absence.

6. Why We Don't Notice Our Computational Processes:

Noticing anything requires a psychological process that starts from some input and leads to a new representation—the representation of the thing that is noticed.

If you would notice something about Process A, your noticing would be the product of Process B. And if you noticed something about Process B, that would require Process C, and so on.

Conclusion: Not every mental process can be noticed. Most mental processes must be carried out without us noticing how they are carried out.

We should expect severe limits on how many things we can notice in our own minds at any given time.

7. Noticing and Absence:

By definition, when you don't notice something, you're not noticing it is an absence.

Noticing an absence has no specific feeling, because not noticing isn't a "thing"—it's the absence of a thing.

Therefore, the fact that we don't notice computational processes in vision, memory retrieval, language processing, etc., is not surprising—we simply aren't representing those processes.

8. Feeling Free in Voluntary Actions:

When we feel unfree, we feel like our bodies are controlled, or we're fighting strong urges, or we can't make ourselves do something we want to do.

Ordinary voluntary mental processes don't feel this way. Instead, they feel free.

Thus, the feeling of being free tells us nothing about whether the underlying process is computational. It simply tells us that the process is ordinary and untroubled.

Summary:

Today, we discussed two arguments suggesting that conscious experience shows mental processes are not computational. But there might be other arguments focusing on how it feels to be a conscious being who thinks, remembers, understands, sees, hears, and feels.

Many people reject the idea that mental processes are computations. Why?

- **It doesn't seem consistent with experience.**
- **It doesn't seem like a clear hypothesis.**

What We Need from a Theory of Computation:

- **A good theory of ordinary computing.**
- It must not rely on any psychological states or processes.
- It must make it possible to explain how brains could perform computations.

Features of Computations:

- Computations are processes that take time.
- They have inputs that are transformed into outputs.
- Both inputs and outputs are **representations**.
- There is a stable pattern in this transformation.

Computation in Arithmetic:

Unfortunately, we don't have theories of computation that are as well-developed as our theories of representation. Today, we'll explore a few ideas, but this remains an area for future research.

One simple idea:

- Computing a function is just a process that starts with an input value(s) and ends with the output value for that function.

But it’s not that easy. Why?

Flashy Numbers Are Not Computation:

- Flashing the numbers "2", "2", and "4" in sequence is not the same as computing the sum of 2 and 2.

Maybe the issue is complexity?

- Flashing numbers isn’t complex enough to match the richness of arithmetic.

But complexity isn’t the only factor:

Cash Register vs. Addition Table:

- An old-school mechanical cash register **actually computes** by adding and subtracting, while a large addition table **does not**, despite being similarly complex.
- Why is this the case?
    - A cash register can compute sums; an addition table just shows them. The register **computes**, while the table simply displays pre-calculated results.

What Makes a Process Computational?

- It's not just about movement or change—a film of numbers wouldn’t count as computation either.

Does input play a special role?

- Imagine inputting numbers into an addition table where a light moves along rows and columns. Is that computation?

Let’s think about it differently:

Everyday Math and Memory:

When you’re figuring out how many avocados you have before a party—six in the fridge and five on the counter, so eleven—are you computing?

- **Not really.** You’re likely recalling basic addition facts that you memorized as a child. You’re **remembering** the result rather than computing it.

Bizarre Result:

- This brings us to an odd conclusion: **grade school math** and **ordinary thinking** don’t actually involve computation, yet an old cash register does?

Computation as Causation Through Abstract Relations:

Maybe it would help to think about the **nature of the addition function**:

- It’s like climbing up a number line one step at a time.
- A mechanical cash register gives “4” from an input of “2 + 2” because of its rotating inner wheels, which mimic this number line.
- **The representation of “4” is reached** because it is where you land when you step up twice from “2” on a number line.
- Thus, the representation of "4" is reached **because it is the sum of 2 and 2.**

Key Idea for Understanding Computation:

- Maybe this abstract relation—between inputs, outputs, and the way they are connected (like steps on a number line)—is key to understanding computation.

This idea is still at the cutting edge of research and needs further development.

Levels of Explanation

Our approach to understanding the mind is to see it as a system of **representations** and **computations** over those representations. However, there is an additional layer of complexity we need to address: **levels of explanation**.

Our guide to understanding these levels is **David Marr**.

Who Was David Marr?

- David Marr was a theoretical psychologist who worked on a variety of topics, including vision.

- He realized that the problem of vision was too complex to solve with tools from just one discipline.
- Marr combined insights from computer science, psychology, neuroscience, and even philosophy to develop a comprehensive explanation of vision.
- In doing so, he created a **systematic theory of psychological explanations** to justify his approach.

Marr’s Three Levels of Explanation:

Marr argued that we need three harmonizing levels of explanation for any mental phenomenon to fully understand it:

1. **Computation**
2. **Representation/Algorithm**
3. **Implementation**

- The **computational level** is the most abstract and general.
- The **algorithmic (and representational) level** is intermediate.
- The **implementation level** is the most concrete and specific.

Today, we’ll focus on the **computational level**.

The Computational Level:

A computational explanation of a mental process answers three key questions:

1. **What are the inputs?**
2. **What are the outputs?**
3. **What is the functional relationship between them?**

This is a precise way of describing **what** mental process we are discussing. Marr believed that if we can’t state the nature of a mental process at this level of precision, we don’t truly understand it, and thus a good explanation would be impossible.

Example: Falling in Love

Let’s consider **falling in love** as a psychological process:

- You meet someone, experience some combination of physical, emotional, and intellectual attraction, spend time together, and maybe fall in love.

While this might describe the process, it doesn't explain it well.

- **What mental states are involved in falling in love?** (inputs)
- **Is falling in love a change in belief, desire, or emotion?** (outputs)
- **What influences the rate or strength of falling in love?** (the functional relationship)

Marr argues that a correct **computational-level explanation** of falling in love would answer these questions. If we can't, then we don't have a good explanation of the process.

A Philosophical Theory of Love:

One theory suggests that for **A to love B** is for **A to desire what is best for B** (for B's own sake). Does this give us a computational theory of love?

- It provides part of the answer:
  - **Input:** unknown
  - **Computation:** unknown
  - **Output:** a desire that B has what is best for them, an imperative representation of the desire type.

While it gives a piece of the explanation, it doesn't yet offer a full computational theory of love.

Easier Example: A Cash Register

Let's switch to something simpler: a basic addition operation in an old-fashioned cash register, like adding **17 + 14** and getting **31** as the result.

For this, the computational level of explanation asks:

- **What are the inputs?**
  - 17, 14, and the addition button.
- **What is the output?**
  - 31.
- **What function is computed?**
  - Addition.

Why Is Love Harder to Explain?

While it's easy to answer these questions for an addition operation in a cash register, it’s much harder to answer them for love because love is more complicated. The difficulty doesn't come from the concept of computational explanation itself but from the **complexity of love** as a mental process.

Final Thought: Does the Cash Register Compute?

Does the cash register actually compute the addition function? Why or why not?

This leads to deeper questions about what we consider computation, but Marr's levels of explanation help clarify what we’re trying to understand about both simple and complex processes.

Marr’s Three Levels of Explanation

Marr says we need **three harmonizing explanations** of any mental phenomenon in order to understand it:

1. **Computation**
2. **Representation/Algorithm**
3. **Implementation**

The Middle Level: Representation and Algorithm

Now, let's focus on the **middle level**: explanation in terms of **representation and algorithm**.

- **How are the inputs and outputs represented?**
- **What algorithm is used to compute the function** that transforms inputs into outputs?

An explanation in terms of **representation and algorithm** tells us **how** inputs are transformed into outputs according to a certain function. It's easy to overlook this because, when we write down a function, we typically use a representational system that also implies the algorithm we would use.

Example: Alphabetizing Names and Phone Numbers

Suppose you want to organize a list of names and phone numbers alphabetically by name. Here’s what the computational process looks like:

- **Input:** An unorganized list of pairs in the form `<name, number>`.
- **Output:** An organized list of `<name, number>` pairs in alphabetical order.
- **Computation:** The process of **alphabetization**.

But what about the **algorithm** to carry out this task?

You could do it manually, writing each name in Roman characters and each number in Arabic numerals, placing the `<name, number>` pairs on index cards, and comparing the letters in the surnames to sort them. Or, you could use software, like Excel, or even code a solution in SQL or C++.

Algorithmic Explanation of Falling in Love?

In a previous discussion, we talked about the process of **falling in love**. It would be nice to describe the algorithm that computes this decision, but we don't even know what the computation is—so we definitely can't determine the algorithm!

The Cash Register Example Revisited

Let's return to the simpler example of a cash register performing **addition**. To understand it at the algorithmic level, we ask:

1. **How are the inputs and outputs represented?**
   - The numbers are represented in a **digital base-10 system**, allowing values from 0.00 to 99.99.
   - Many other ways of representing numbers could have been used, but this system was chosen.
2. **What algorithm is used to carry out the addition?**
   - Add the 10^{-2} values. If the result is greater than 9, add 1 to the 10^{-1} value.
   - Add the 10^{-1} values, and so on through 10^{0}, 10^{1}, and 10^{2}.
   - Stop when all values have been added and recorded.

The Bottom Level: Implementation

The final level of explanation is **implementation**, which asks: **How do physical objects embody the representational system and execute the algorithm?**

In the case of an old-fashioned cash register, implementation is straightforward:

- **Representation:** The input values are represented by the positions of wheels. For example, the first wheel's position represents 0.00-0.09, the second wheel represents 0.0-0.9, and so on.
- **Execution:** The physical positions of the wheels change as they interact. For instance, when the wheel representing 10^{-2} rotates from 9 to 0, a catch moves the 10^{-1} wheel up by one position.

For mechanical systems, the level of implementation is relatively easy to grasp.

Challenges with Neural Implementation

However, moving from a **psychological** explanation to a **neural** one is much trickier. How do neurons and their connections embody representations or follow algorithms? For example:

- **How is a representation made out of neurons?**
- **How is a computation performed by neurons through chemical signaling?**

For something like **love**, we don't know the computational process or the precise algorithm. But we do know some aspects of the **neural implementation**. For instance, **fMRI imaging** allows us to observe which brain regions are more active during specific tasks, which may give clues about where computations and representations related to love occur.

The Complexity of fMRI and Brain Activity

There are, however, difficulties with this approach:

- **What does fMRI activity represent?** It's unclear whether it measures input, local activity, output, or some combination of these.
- **Confounding factors:** There are often poorly defined categories of investigation and difficulties in pinpointing what fMRI activity actually reflects.

Despite these challenges, many philosophers and scientists believe we have some crude ideas about how love is implemented in the brain—even though we lack clarity on the exact computational process.

Conclusion: Levels of Explanation Are Not Purely Top-Down

Understanding the mind doesn't have to start from the top-down, beginning with abstract computational models. You can begin anywhere—whether with rough guesses or partial information—at any level of explanation.