Because they only articulate the meaning of one sentence in terms of another sentence, they do not depend on experience in any way and so cannot be refuted by experience. Such statements are often referred to as *analytic statements* to distinguish them from ordinary sentences whose truth depends upon how the world is (*synthetic statements*)

The criterion that theoretical terms have to be translatable into observational terms was quickly recognized to be too strong

- A number of theoretical terms i.e. dispositional terms like *soluble* may not be translatable into observational terms (many soluble objects will never be placed into water)
- The dispositional term cannot be translated into a conditional sentence (e.g. if placed into water then it will dissolve)
    - The reason being that in symbolic logic a sentence of the form "if--, then..." us defined as true if the *antecedent* is false == making any object never placed in water soluble

To account for the meaning of such terms, the positivists weakened their verifiability condition

- Carnap proposed that a *dispositional* term like soluble could be translated by the following sentence (a *reduction sentence*)

$$\text{"if x is placed in water, then x will dissolve if and only if x is soluble"}$$

The Deductive-Nomological Model of Explanation and the Hypothethico-Deductive Model of Theory Development

LP wanted to predict and explain phenomena. The positivists maintained that explaining an event consisted of deriving a statement describing antecedently known empirical facts (initial conditions). Thus deduction plays a central role in their account of explanation and the Positivists adopted what has been termed the "covering law" or "deductive-nomological" (D-N) model of scientific explanation.
This basic model is represented by the following schema in which L_1 through L_n represent general laws, C_1 through C_n represent initial conditions and E represents the event to be explained:

$$L_1, L_2, \ldots L_n$$
$$C_1, C_2, \ldots C_n$$

A couple of features of this general schema of explanation should be noted.

1. in order to explain an event according to LP, it was not sufficient to simply point to a factor that might have caused the event (thrown rock caused broken window)
    1. An explanation requires a complete derivation of the event from general laws and known facts
2. There is a symmetry between explanation and prediction. They have the same logical structure such that a derivation of the sort required for explanation would if carried out before the event serve to predict the event

Several Positivists tried to modify this model to include probabilistic laws in addition to the deterministic ones. For instance, that taking a particular drug may cure a disease most of but not all the time

- Hempel (1962) proposed a modification of the D-N model to allow for "inductive-statistical" explanations (only for event above the .5 statistical regularity)

The Positivists called the procedure for developing scientific laws the **hypothetico-deductive*** method (H-D)

- The basic idea of H-D method is that scientists begin with an event that requires explanation

Both models however encounter problems which were recognized by the Positivists themselves

- The D-N model requires that one of the premises in the deductive explanation of an event be a law
    - Explicating what makes a statement into a law, however, is a difficult problem given the tools of symbolic logic
    - It was clear that a law statement had to be a true general statement of the form: "For all x, if x is F, then x is G" or $[(x)(Fx \rightarrow Gx)]$
        - "for any person, if the person is infected with cadaveric matter, then the person contracts childbed fever"
    - However it was also clear that this is insufficient because we would not want to count all true general statements as laws
    - It is commonly thought that laws are more than general statements that happen to be true -- We think they tell us something about the limits of how things **must be**

"[...] [...] What is important to note here is that being explained by a theory is the only factor to which the Positivists can appeal to distinguish laws from universal statements"

The use of symbolic logic also poses problems for the H-D analysis of hypothesis development

- Hume recognized that inductive evidence could never establish definitively the truth of any general claim
  - Undiscovered counterevidence is always possible
- A number of paradoxes were brought forward to challenge the assumption that confirming evidence should strengthen our belief in particular hypotheses
  - Raven Paradox: The *raven paradox* depends on the fact that a law statement of the form:
    - "For all x, if x is F, then x is G (see above) is logically equivalent to the statement "For all x, if x is not G, then x is not F"
    - Basically "for all x, if x is a raven then it is black" is logically equivalent to == "all things that are not black are not ravens"
      - Following the H-D model the form which it is logically equivalent to only requires us to examine things that are not black and test the prediction that these things will not be ravens

> !! "The Positivists' commitment to symbolic logic and, in particular, their commitment to laws being fundamentally universal generalizations, however, lay at the heart of these problems. Hence, they were not easily resolved. Moreover, the moves made to rescue Positivism tend to cloud the clear and intuitive picture of the nature of explanation and confirmation that the Positivists account seemed to offer" Bechtel, p. 27 (6 in pdf)

The Axiomatic Account of Theories

For the Logical Positivists a theory was a structured network of statements from which one could derive specific laws, e.g. Euclidean geometry

- At the core of this are a set of primitive terms and postulates from which various axioms can be derived
- In a like manner Positivists proposed scientific theories could themselves be rendered as deductive structures --> set of primitive terms and postulates

    - The particular laws would be the axioms that we could derive from these assumptions

The LPs also envisioned that the process of axiomatizing theories could bring unity to science

- The text gives example of astronomy being swallowed up by physics because Newton was a boss who figured shit out
    - "Newton showed that the basic postulates of astronomical theory are themselves axioms derived from a more basic physical theory" --> Astronomy was thus subsumed within physics
- Eventually the Positivists proposed, all sciences could be subsumed into one theoretical edifice, that of unified science
    - The process of unifying science by deriving principles of one science from those of another is commonly spoken of as *theory reduction*
        1. This assumes that all science is a cumulative enterprise --> larger and larger theoretical networks
        2. It views the laws of specialized disciplines as derivative laws which in principle can be derived from physics

Summary of Logical Positivism

The Logical Positivists offered:

- A systematic and highly attractive view of the project of science
- A theory of meaning that showed how scientific discourse was grounded in sensory experience and thus certain to be meaningful
- An account of explanation that used deduction to show how particular events could be explained by laws and an account of confirmation that showed how particular events provided evidence for the laws that were developed
- An explanation of how the laws of each science could be unified in to axiomatic structures and ultimately grounded in a unified account of nature

How might these doctrines apply to cognitive science?

- Some doctrines had a significant effect on behaviorists such as Spence and Skinner
    - Verificationist theory of meaning

Although less popular during the reign of cognitivism one can illustrate the basic claims of the Positivists equally by showing how they could be applied to theories of recent cognitive science

- Structure of human concepts and the processes of categorization (see Knowledge +Conceptual Knowledge)
- To develop a scientific theory of concepts the Ps would insist it is necessary first to provide a criteria for the meaning of terms used in the theorizing especially *concept*
    - Rosch (1975, 1978) experiments on typicality
        - Meaning & Degree of Category Membership
- P's wants science to explain and predict phenomena in nature via laws
    - One such in Cognitive science is prototype theory which has been corroborated

Challenges to Logical Positivism

The Attack on Confirmation

Karl Popper challenged especially the hypothetico-deductive model

- Just as Hume showed that it was impossible to prove that a general statement was true, Popper maintained that one could not even show it was likely to be true
- Popper proposed a radical remedy to this problem: He recommended abandoning the whole endeavor of seeking well-confirmed theories and proposed instead a focus on falsifying hypotheses
    - *Modus tollens* or denying the consequent: Modus Tollens
    - Popper seeks **corroboration** not confirmation
    - The more things a theory rules out the more powerful the theory

Informative $\iff$ *Unlikely*

Falsificationism - Lecture

Learning Goals

1. Understanding the induction problem
2. Understanding how falsification follows deductive procedures
3. Understanding how Popper understands falsifiability as characteristic that separates scientific from non-scientific statements (solves the demarcation problem)

Reading tips for next time

Puzzle solving
paradigms
revolution
normal science
incommensurability

NU SKAL DU FICKIGN HÆRE JAMMERAT

Falsificationism: The view that scientific statements are characterized by there being procedures that *could* show the statements to be false

In a nut shell:
Statement: "All ravens are black"

R: is a raven
B: is black
$R \rightarrow B$: if something is a raven, then it is black
, : separating premises
$x \vdash y$: y's truth follows from x's truth
$\neg B$: B is false

$R \rightarrow B, \neg B \vdash \neg R$
Modus tollens

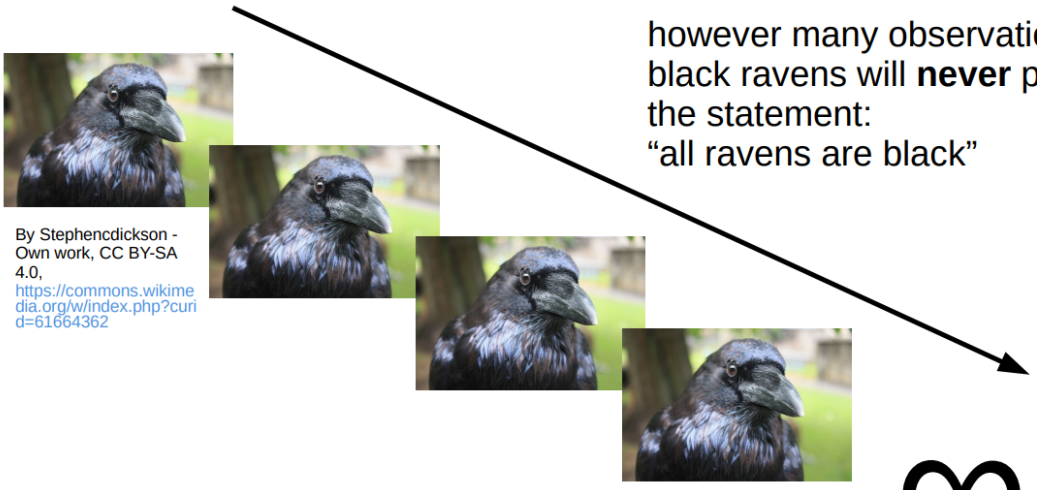
Valid arguments: the truth of the premises guarantee the truth of the conclusion

T: a theory
O: an observation
$T \rightarrow O$: T entails O

$T \rightarrow O, O \vdash T$: is invalid = Induction problem
$T \rightarrow O, \neg O \vdash \neg T$: is valid = A single observation of a white raven will disprove the

statement

- Thus, observations, O, that conform to our theory, T, do not confirm the truth of T (Induction problem)
- Though, observations, O, that do not conform to our theory, T, falsify T.



$T \rightarrow O,\ O \vdash T$: is invalid

T	$\rightarrow$	O
T	T	T
T	F	F
F	T	T
F	T	F

$T \rightarrow O, O \vdash T$:
is similar to
$((T \rightarrow O) \wedge O) \rightarrow T$
"A complicated proposition" - you're not saying anything about validity

Although in the first one, we're making the claim that the form is valid

- This is a truth statement

Validity $\neq$ meaning
Valid statements can be useless
i.e.
$T \wedge \neg T \vdash O$ is a valid statement but does not say anything

The Backdrop

Logical positivism

- influential from the 1920s
- some names: Rudolf Carnap, Herbert Feigl, Hans Reichenbach, Carl Hempel
- emphasized experience as the way to build knowledge
- also tried to create a formal (logical) language to express scientific statements in
- espoused the verifiability theory of meaning
- espoused the deductive-nomological model of explanation

Induction Problem

- Induction: passing from singular statements (single ravens) to universal statements (all ravens)
- We cannot prove by induction, i.e. truth of conclusion doesn't follow from truth of premises
  - a weak version may be that we provide evidence
- Seemingly irrelevant evidence must be admitted as evidence (raven paradox)

Deduction

- Truth of conclusion follows from truth of premises
- But we cannot confirm universal statements with deduction
  -$R \rightarrow B, B \vdash R$ is a fallacy (confirming the antecedent)
- Popper bites the bullet and makes do with falsifying statements
- We can go from singular statements (a white raven) to negation of universal statements (all ravens are black)

The Problem of Demarcation

- separating science from non-science

> "The problem of finding a criterion which would enable us to distinguish between the empirical sciences on the one hand, and mathematics and logic as well as 'metaphysical' systems on the other, I call the problem of demarcation" (Popper 2002, p. 11)

Principle of demarcation

- "It must be possible for an empirical scientific system to be refuted by experience" (Popper 2002, p. 18)

Attack on Psychoanalysis

- Psychoanalysis
  - Only explanations are given, no predictions provided
  - Everything is compatible with, say, childhood trauma, repression, inferiority feelings, etc.
    - Makes falsification impossible

Normative View of Scientific Practice

- State a theory, T
  - the “new theory should proceed from some simple, new, and powerful, unifying idea about some connection or relation” (Popper 1965)
- Operationalize those theories, i.e. make predictions, $T \rightarrow O$
  - independent, testable implications that have not been tested before
- Do your best to **falsify** T, i.e. try as hard as you can to find $\neg O$
- So we can’t confirm T, but we can **corroborate** T by trying our best to falsify it

Duhem's Thesis

"*Typically, T does not deductively imply O; rather, it is T&A that deductively implies O (here, T is a theory, O is an observation statement, and A is a set of of auxiliary assumptions)*"

importantly, the auxiliary assumptions should be independently testable

Auxiliary theories

- Theories *T*, do not exist in isolation
  - They are supported by auxiliary theories, $[\mathrm{A}_1, \ldots, \mathrm{A}_\mathrm{n}]$

$(A_1 \wedge T) \rightarrow O, \neg O \vdash \neg T$; is this valid?
$(A_1 \wedge T) \rightarrow O, \neg O \vdash \neg(A_1 \wedge T)$; is this valid?

In cognitive science everything is interdependent

- Memory depends on perception, which depends on sensation

Popper - A Survey of Some Fundamental Problems

THE PROBLEM OF INDUCTION

It is usual to call an inference ‘inductive’ if it passes from *singular statements* (sometimes also called ‘particular’ statements), such as accounts of the results of observations or experiments, to *universal statements*, such as hypotheses or theories

> !! "Now it is far from obvious, from a logical point of view, that we are justified in inferring universal statements from singular ones, no matter how numerous; for any conclusion drawn in this way may always turn out to be false: no matter how many instances of white swans we may have observed, this does not justify the conclusion that *all* swans are white." p. 4

The question whether inductive inferences are justified, or under what conditions, is known as the **problem of induction**

Yet if we want to find a way of justifying inductive inferences, we must first of all try to establish a principle of induction. A **principle of induction** would be a statement with the help of which we could put inductive inferences into a logically acceptable form

> !! "Now this principle of induction cannot be a purely logical truth like a tautology or an analytic statement. Indeed, if there were such a thing as a purely logical principle of induction, there would be no problem of induction [...] Thus the principle of induction must be a synthetic statement; that is, a statement whose negation is not self-contradictory but logically possible. So the question arises why such a principle should be accepted at all, and how we can justify its acceptance on rational grounds."

In short, like every other form of inductive logic, the logic of probable inference, or 'probability logic', leads either to an infinite regress, or to the doctrine of apriorism.

"The theory to be developed in the following pages stands directly opposed to all attempts to operate with the ideas of inductive logic. It might be described as the theory of the deductive method of testing, or as the view that a hypothesis can only be empirically tested—and only after it has been advanced. "

ELIMINATION OF PSYCHOLOGISM

Justification or validity (Kant's quid juris?)
Its questions are of the following kind.

- Can a statement be justified?
- And if so, how? Is it testable?
- Is it logically dependent on certain other statements?
- Or does it perhaps contradict them?

DEDUCTIVE TESTING OF THEORIES

From a new idea, put up tentatively, and not yet justified in any way—an anticipation, a hypothesis, a theoretical system, or what you will—conclusions are drawn by means of logical deduction. These conclusions are then compared with one another and with other relevant statements, so as to find what logical relations (such as equivalence, derivability, compatibility, or incompatibility) exist between them.

We may if we like distinguish four different lines along which the testing of a theory could be carried out.

1. there is the **logical comparison** of the conclusions among themselves, by which the internal consistency of the system is tested.
2. there is the investigation of the logical form of the theory, with the object of determining whether it has the character of an **empirical** or **scientific theory**, or whether it is, for example, **tautological**.
3. there is the comparison with other theories, chiefly with the aim of **determining whether the theory would constitute a scientific advance** should it survive our various tests.
4. finally, there is the testing of the theory by way of **empirical applications of the conclusions** which can be derived from it.

It should be noticed that a positive decision can only temporarily *support* the theory, for subsequent negative decisions may always overthrow it. So long as theory withstands detailed and severe tests and is not superseded by another theory in the course of scientific progress, we may say that it has 'proved its mettle' or that it is 'corroborated' by past experience

THE PROBLEM OF DEMARCATION

In rejecting the method of induction, it may be said, I deprive empirical science of what appears to be its most important characteristic; and this means that I remove the barriers which separate science from metaphysical speculation. My reply to this objection is that my main reason for rejecting inductive logic is precisely that it *does not provide a suitable distinguishing mark* of the empirical, non-metaphysical, character of a theoretical system; or in other words, that it *does not provide a suitable 'criterion of demarcation'*.

This problem was known to Hume who attempted to solve it. With Kant it became the central problem of the theory of knowledge. If, following Kant, we call the problem of induction 'Hume's problem', we might call the problem of demarcation 'Kant's problem'.

The older positivists wished to admit, as scientific or legitimate, only those *concepts* (or notions or ideas) which were, as they put it, 'derived from experience'; those concepts, that is, which they believed to be logically reducible to elements of sense-experience, such as sensations (or sense-data), impressions, perceptions, visual or auditory memories, and so forth

(Logical Positivism)

Modern positivists are apt to see more clearly that science is not a system of concepts but rather a system of *statements*. Accordingly, they wish to admit, as scientific or legitimate, only those statements which are reducible to elementary (or 'atomic') statements of experience—to 'judgments of perception' or 'atomic propositions' or 'protocol-sentences' or what not

- It is clear that the implied criterion of demarcation is identical with the demand for an inductive logic

> "Since I reject inductive logic I must also reject all these attempts to solve the problem of demarcation. With this rejection, the problem of demarcation gains in importance for the present inquiry. Finding an acceptable criterion

> of demarcation must be a crucial task for any epistemology which does not accept inductive logic"

Positivists usually interpret the problem of demarcation in a *naturalistic way*; they interpret it as if it were a problem of natural science. [...] They are constantly trying to prove that metaphysics by its very nature is nothing but nonsensical twaddle—'sophistry and illusion', as Hume says, which we should 'commit to the flames'

Popper criticizes positivism some more:
"[...] we find that each time the positivists tried to say more clearly what 'meaningful' meant, the attempt led to the same result—to a definition of 'meaningful sentence' (in contradistinction to 'meaningless pseudo-sentence') which simply reiterated the criterion of demarcation of their *inductive logic.*"

"This 'shows itself' very clearly in the case of Wittgenstein, according to whom every meaningful proposition must be *logically reducible* to elementary (or atomic) propositions, which he characterizes as descriptions or 'pictures of reality'5"

" And it is precisely over the problem of induction that this attempt to solve the problem of demarcation comes to grief: positivists, in their anxiety to annihilate metaphysics, annihilate natural science along with it. For scientific laws, too, cannot be logically reduced to elementary statements of experience. If consistently applied, Wittgenstein's criterion of meaningfulness rejects as meaningless those natural laws the search for which, as Einstein says, is 'the supreme task of the physicist': they can never be accepted as genuine or legitimate statements"

Wittgenstein's attempt to unmask the problem of induction as an empty pseudo-problem was formulated by Schlick in the following words: ' The problem of induction consists in asking for a logical justification of universal statements about reality . . . We recognize, with Hume, that there is no such logical justification: there can be none, simply because they are not genuine statements .'

This shows how the inductivist criterion of demarcation fails to draw a dividing line between scientific and metaphysical systems, and why it must accord them equal status; for the verdict of the positivist dogma of meaning is that both are systems of meaningless pseudostatements. Thus instead of eradicating metaphysics from the empirical sciences, positivism leads to the invasion of metaphysics into the scientific realm

My criterion of demarcation will accordingly have to be regarded as a *proposal for an agreement or convention*

Popper doesn't hate metaphysics!!!!

"For it cannot be denied that along with metaphysical ideas which have obstructed the advance of science there have been others—such as speculative atomism—which have aided it. "

Then backs down a bit lol

"Yet having issued all these warnings, I still take it to be the first task of the logic of knowledge to put forward a *concept of empirical science*, [...] in order to draw a clear line of demarcation between science and metaphysical ideas—even though these ideas may have furthered the advance of science throughout its history"

EXPERIENCE AS A METHOD

We may distinguish three requirements which our empirical theoretical system will have to satisfy:

1. First, it must be **synthetic**, so that it may represent a non-contradictory, a **possible** world
2. Secondly, it must satisfy the criterion of demarcation, i.e. it must not be metaphysical, but must represent a world of possible **experience**.
3. Thirdly, it must be a system distinguished in some way from other such systems as the one which represents **our** world of experience

But how is the system that represents our world of experience to be distinguished? The answer is: by the fact that it has been submitted to tests, and has stood up to tests. This means that it is to be distinguished by applying to it that deductive method

FALSIFIABILITY AS A CRITERION OF DEMARCATION

Now in my view there is no such thing as induction. Thus inference to theories, from singular statements which are ‘verified by experience’ (whatever that may mean), is logically inadmissible. Theories are, therefore, never empirically verifiable. If we wish to avoid the positivist’s mistake of eliminating, by our criterion of demarcation, the theoretical systems of natural science, then we must choose a criterion which allows us to admit to the domain of empirical science even statements which cannot be verified

> "But I shall certainly admit a system as empirical or scientific only if it is capable of being tested by experience. These considerations suggest that

> not the verifiability but the falsifiability of a system is to be taken as a criterion of demarcation."
> "In other words: I shall not require of a scientific system that it shall be capable of being singled out, once and for all, in a positive sense; but I shall require that its logical form shall be such that it can be singled out, by means of empirical tests, in a negative sense: it must be possible for an empirical scientific system to be refuted by experience.[1]"

since the amount of positive information about the world which is conveyed by a scientific statement is the greater the more likely it is to clash, because of its logical character, with possible singular statements. (Not for nothing do we call the laws of nature ‘laws’: the more they prohibit the more they say.)

My proposal is based upon an asymmetry between verifiability and falsifiability; an *asymmetry* which results from the logical form of universal statements:

- For these are never derivable from singular statements, but can be contradicted by singular statements.
  - Consequently it is possible by means of purely deductive inferences (with the help of the modus tollens of classical logic) to argue from the truth of singular statements to the falsity of universal statements
  - Such an argument to the falsity of universal statements is the only strictly deductive kind of inference that proceeds, as it were, in the ‘inductive direction’; that is, from singular to universal statements.
- BASICALLY Popper is saying that only specific instances can falsify universal statements

A third objection may seem more serious. It might be said that even if the asymmetry is admitted, it is still impossible, for various reasons, that any theoretical system should ever be conclusively falsified.

- For it is always possible to find some way of evading falsification, for example by introducing ad hoc an auxiliary hypothesis, or by changing ad hoc a definition.
- It is even possible without logical inconsistency to adopt the position of simply refusing to acknowledge any falsifying experience whatsoever

According to my proposal, what characterizes the empirical method is its manner of exposing to falsification, in every conceivable way, the system to be tested.

- Its aim is not to save the lives of untenable systems but, on the contrary, to select the one which is by comparison the fittest, by exposing them all to the fiercest

struggle for survival.

THE PROBLEM OF THE ‘EMPIRICAL BASIS’

Problems of the empirical basis—that is, problems concerning the empirical character of singular statements, and how they are tested—thus play a part within the logic of science that differs somewhat from that played by most of the other problems which will concern us.

For most of these stand in close relation to the **practice** of research, whilst the problem of the empirical basis belongs almost exclusively to the **theory** of knowledge. I shall have to deal with them, however, since they have given rise to many obscurities. This is especially true of the relation between **perceptual experiences** and **basic statements**. (What I call a ‘basic statement’ or a ‘basic proposition’ is *a statement which can serve as a premise in an empirical falsification; in brief, a statement of a singular fact.*)

We must distinguish between, on the one hand, our **subjective experiences** or our **feelings of conviction**, which can never justify any statement (though they can be made the subject of psychological investigation) and, on the other hand, the **objective logical relations** subsisting among the various systems of scientific statements, and within each of them

SCIENTIFIC OBJECTIVITY AND SUBJECTIVE CONVICTION

> "The words ‘objective’ and ‘subjective’ are philosophical terms heavily burdened with a heritage of contradictory usages and of inconclusive and interminable discussions."

My use of the terms ‘objective’ and ‘subjective’ is not unlike Kant’s. He uses the word ‘**objective**’ to indicate that scientific knowledge should be *justifiable*, independently of anybody’s whim: a justification is ‘objective’ if in principle it can be tested and understood by anybody. ‘If something is valid’, he writes, ‘for anybody in possession of his reason, then its grounds are objective and sufficient.'

Now I hold that scientific theories are never fully justifiable or verifiable, but that they are nevertheless testable. I shall therefore say that the objectivity of scientific statements lies in the fact that they can be inter-subjectively tested [2]

The word '**subjective**' is applied by Kant to our feelings of conviction (of varying degrees).

Kant was perhaps the first to realize that the objectivity of scientific statements is closely connected with the construction of theories— with the use of hypotheses and universal statements. Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested—in principle—by anyone.

Popper completely disregards that subjective conviction of some fact serves as evidence of anything, which makes great sense

In demanding objectivity for basic statements as well as for other scientific statements, we deprive ourselves of any logical means by which we might have hoped to reduce the truth of scientific statements to our experiences.

We thus arrive at the following view. Systems of theories are tested by deducing from them statements of a lesser level of universality. These statements in their turn, since they are to be inter-subjectively testable, must be testable in like manner—and so ad infinitum.

> "It might be thought that this view leads to an infinite regress, and that it is therefore untenable. [...] The deductive method of testing cannot establish or justify the statements which are being tested; nor is it intended to do so. Thus there is no danger of an infinite regress."

> "I do not demand that every scientific statement must have in fact been tested before it is accepted. I only demand that every such statement must be **capable** of being tested; or in other words, I refuse to accept the view that there are statements in science which we have, resignedly, to accept as true merely because it does not seem possible, for logical reasons, to test them."

1. Note that I suggest falsifiability as a criterion of demarcation, but not of meaning. Falsifiability separates two kinds of perfectly meaningful statements: the falsifiable and the non-falsifiable. It draws a line inside meaningful language, not around it↩
2. I have since generalized this formulation; for inter-subjective testing is merely a very important aspect of the more general idea of inter-subjective criticism, or in other words, of the idea of mutual rational control by critical discussion.↩

Popper - Theories

The empirical sciences are systems of theories. The logic of scientific knowledge can therefore be described as a theory of theories

CAUSALITY, EXPLANATION, AND THE DEDUCTION OF PREDICTIONS

To give a causal explanation of an event means to deduce a statement which describes it, using as premises of the deduction one or more universal laws, together with certain singular statements, the initial conditions.

We have thus two different kinds of statement, both of which are necessary ingredients of a complete causal explanation. They are

1. universal statements, i.e. hypotheses of the character of natural laws, and
2. singular statements, which apply to the specific event in question and which I shall call ‘initial conditions’.

It is from universal statements in conjunction with initial conditions that we deduce the singular statement,

- We call this statement a specific or singular prediction

The ‘principle of causality’ is the assertion that any event whatsoever can be causally explained—that it can be deductively predicted. According to the way in which one interprets the word ‘can’ in this assertion, it will be either tautological (analytic), or else an assertion about reality (synthetic).

But in this case it (principle of causality) is not falsifiable, as will be seen later, in section 78. I shall, therefore, neither adopt nor reject the ‘principle of causality’; I shall be content simply to exclude it, as ‘metaphysical’, from the sphere of science.

I shall, however, propose a methodological rule which corresponds so closely to the ‘principle of causality’ that the latter might be regarded as its metaphysical version.

- It is the simple rule that we are not to abandon the search for universal laws and for a coherent theoretical system, nor ever give up our attempts to explain causally any kind of event we can describe

STRICT AND NUMERICAL UNIVERSALITY

We can distinguish two kinds of *universal synthetic statement*: the '**strictly universal**' and the '**numerically universal**'. It is the strictly universal statements which I have had in mind so far when speaking of universal statements—of theories or natural laws. The other kind, the numerically universal statements, are in fact equivalent to certain singular statements, or to conjunctions of singular statements, and they will be classed as singular statements here

I consider it both useful and fruitful to regard natural laws as synthetic and strictly universal statements ('all-statements'). This is to regard them as non-verifiable statements which can be put in the form: 'Of all points in space and time (or in all regions of space and time) it is true that . . .'.

- By contrast, statements which relate only to certain finite regions of space and time I call 'specific' or 'singular' statements.

UNIVERSAL CONCEPTS AND INDIVIDUAL CONCEPTS

The distinction between universal and singular statements is closely connected with that between universal and individual concepts or names. It is usual to elucidate this distinction with the help of examples of the following kind: 'dictator', 'planet', 'H2O' are universal concepts or universal names. 'Napoleon', 'the earth', 'the Atlantic' are singular or individual concepts or names.

- In these examples individual concepts or names appear to be characterized either by being proper names, or by having to be defined by means of proper names, whilst universal concepts or names can be defined without the use of proper names.

'universal concepts' and '*individual concepts*'. If I were asked for definitions I should probably have to say, as above: 'An individual concept is a concept in the definition of which proper names (or equivalent signs) are indispensable. **If any reference to proper names can be completely eliminated, then the concept is a universal concept**

STRICTLY UNIVERSAL AND EXISTENTIAL STATEMENTS

It is not enough, of course, to characterize universal statements as statements in which no individual names occur. If the word ‘raven’ is used as a universal name, then, clearly, the statement ‘all ravens are black’ is a strictly universal statement. But in many other statements such as ‘many ravens are black’ or perhaps ‘some ravens are black’ or ‘there are black ravens’, etc., there also occur only universal names; yet we should certainly not describe such statements as universal

Statements in which only universal names and no individual names occur will here be called ‘strict’ or ‘pure’. Most important among them are the **strictly universal** statements which I have already discussed.

- In addition to these, I am especially interested in statements of the form ‘there are black ravens’, which may be taken to mean the same as ‘there exists at least one black raven’. Such statements will be called **strictly or purely existential statements** (or ‘**there-is’ statements**).

The theories of natural science, and especially what we call natural laws, have the logical form of strictly universal statements; thus they can be expressed in the form of negations of strictly existential statements or, as we may say, in the form of **non-existence statements** (or ‘**thereis-not**’ statements)

In this formulation we see that natural laws might be compared to ‘proscriptions’ or ‘prohibitions’. They do not assert that something exists or is the case; they deny it. They insist on the non-existence of certain things or states of affairs, proscribing or prohibiting, as it were, these things or states of affairs: they rule them out. And it is precisely because they do this that they are **falsifiable**.

Strictly existential statements, by contrast, cannot be falsified. No singular statement (that is to say, no ‘basic statement’, no statement of an observed event) can contradict the existential statement, ‘There are white ravens’. Only a universal statement could do this. On the basis of the criterion of demarcation here adopted I shall therefore have to treat strictly existential statements as non-empirical or ‘metaphysical’.

THEORETICAL SYSTEMS

A theoretical system may be said to be axiomatized if a set of statements, the axioms, has been formulated which satisfies the following four fundamental requirements.

- (a) The system of axioms must be free from contradiction (whether self-contradiction or mutual contradiction). This is equivalent to the demand that not every arbitrarily chosen statement is deducible from it
- (b) The system must be independent, i.e. it must not contain any axiom deducible from the remaining axioms. (In other words, a statement is to be called an axiom only if it is not deducible within the rest of the system.)
    - These two conditions concern the axiom system as such
- (c) sufficient for the deduction of all statements belonging to the theory which is to be axiomatized, and
- (d) necessary, for the same purpose; which means that they should contain no superfluous assumptions.

SOME POSSIBILITIES OF INTERPRETING A SYSTEM OF AXIOMS

I consider two different interpretations of any system of axioms to be admissible. The axioms may be regarded either (i) as conventions, or they may be regarded (ii) as empirical or scientific hypotheses.

(i)
If the axioms are regarded as conventions then they tie down the use or meaning of the fundamental ideas (or primitive terms, or concepts) which the axioms introduce; they determine what can and what cannot be said about these fundamental ideas. Sometimes the axioms are described as 'implicit definitions' of the ideas which they introduce

The interpretation of an axiomatic system as a system of (conventions or) implicit definitions can also be expressed by saying that it amounts to the decision: only models may be admitted as substitutes. But if a model is substituted then the result will be a system of analytic statements (since it will be true by convention). An axiomatic system interpreted in this way cannot therefore be regarded as a system of empirical or scientific hypotheses (in our sense) since it cannot be refuted by the falsification of its consequences; for these too must be analytic.

(ii)
How then, it may be asked, can an axiomatic system be interpreted as a system of empirical or scientific hypotheses? The usual view is that the primitive terms occurring in the axiomatic system are not to be regarded as implicitly defined, but as 'extra-logical constants'. For example, such concepts as 'straight line' and 'point', which occur in every axiom system of geometry, may be interpreted as 'light ray' and 'intersection of

light rays'. In this way, it is thought, the statements of the axiom system become statements about empirical objects, that is to say, synthetic statements

Here I may perhaps add that it is usually possible for the primitive concepts of an axiomatic system such as geometry to be correlated with, or interpreted by, the concepts of another system, e.g. physics.

LEVELS OF UNIVERSALITY. THE *MODUS TOLLENS*

We may distinguish, within a theoretical system, statements belonging to various levels of universality. The statements on the highest level of universality are the axioms; statements on the lower levels can be deduced from them.

[...] the way in which the falsification of a conclusion entails the falsification of the system from which it is derived —is the modus tollens of classical logic. It may be described as follows:

- Let p be a conclusion of a system t of statements which may consist of theories and initial conditions (for the sake of simplicity I will not distinguish between them).
- We may then symbolize the relation of derivability (analytical implication) of p from t by 't → p' which may be read: 'p follows from t'. Assume p to be false, which we may write '$\bar{p}$', to be read 'not-p'. Given the relation of deducibility, t → p, and the assumption $\bar{p}$, we can then infer $\bar{t}$ (read 'not-t'); that is, we regard t as falsified.
- If we denote the conjunction (simultaneous assertion) of two statements by putting a point between the symbols standing for them, we may also write the falsifying inference thus:
  - $((t \rightarrow p).\bar{p}) \rightarrow \bar{t}$, or in words: 'If p is derivable from t, and if p is false, then t also is false'.

By means of this mode of inference we falsify the whole system (the theory as well as the initial conditions) which was required for the deduction of the statement p, i.e. of the falsified statement

Cognitive Science History

Cognitive Science's birth has been dated to September 11, 1956, the second day of a Symposium on Information Theory at MIT.

Cognitive Science is largely a multidisciplinary field especially in its infancy. A great many people have historically contributed to our knowledge of cognition and its

processes, which I will here make a list of:

George Miller Pattern Learning and Memory

- A key contributor to the emergence of cognitive science, psychologist
- "The magical number 7, plus or minus 2

AI

George Boole

- British mathematician
- In his 1854 book, The Laws of Thought, Boole demonstrated that formal operations performed on sets corresponded to logical operators (and, or, not) applied to propositions
  - Boole proposed that these could serve as laws of thought
- "Boolean" gets its name from Boole

Newell and Simon THE GPS Means-Ends Analysis Pattern Learning and Memory

- The Nature of Problem Solving
- Problem solvers baby
- Newell and Simon soon went beyond their initial Logic Theorist program to a General Problem Solver they used in less formal domains, such as solving Tower of Hanoi problems.
  - They developed such concepts as subgoals, heuristics, and satisficing and introduced the production system framework, which employs rules that operate on the contents of working memory when their antecedent conditions are satisfied.
- Also developed further systems such as SOAR and EPAM
- A 1968 book reporting this work also included a seminal chapter by Ross Quillian introducing semantic networks. Conceptual Knowledge Meaning

Psychology

B. F. Skinner The History of Cognitive Psychology Development > Chomsky vs Skinner B. F. Skinner - Behaviorism

- Radical behaviorist

    - Actively opposed positing internal processes and focused on what was observable: describing how behavioral responses changed with contingencies of reinforcement.

Jean Piaget Cognitive Development

- Genetic epistemology
    - How knowledge developed in human organisms. The theory considers how children (and adults) come to know their world.

Sir Frederic Bartlett Schemas

- Particularly famous for his work on the concept of schema and his studies on memory and recall. Bartlett challenged traditional views of memory as a passive storehouse of information and instead emphasized the constructive and dynamic nature of memory.
    - Memory distortion

Donald Broadbent Implicit vs Explicit Memory Broadbent's Filter Theory

- Information processing and memory

Eleanor Rosch Meaning Bechtel - Philosophy of Science

- A Bruner student, began work that led in the 1970s to a view of categories that emphasized prototypes, fuzzy boundaries, and the primacy of basic-level categories.

Atkinson and Shiffrin Encoding and Storage

- In 1968 developed a model that integrated previous work on control processes, sensory memory (Sperling), short-term memory (Peterson and Peterson), and the distinction between short-term and long-term memory (William James; Waugh and Norman).

Roger Shepard Visual Imagery Encoding and Storage Knowledge

- Roger Shepard did elegant work in mathematical psychology (e.g., he pioneered nonparametric multidimensional scaling), but is best known for his research on mental imagery and mental rotation with such students as Lynn Cooper and Jacqueline Metzler.
    - For example, they demonstrated that when subjects had to decide whether a comparison stimulus was a rotation or a mirror image of a geometrical form, their reaction times increased linearly with the degree of rotation

    - This suggested that subjects mentally rotated the comparison stimulus—an attention-grabbing claim at a time when mentalism was still suspect in many quarters

Gordon Bower

- Moved from mathematical models of learning towards more cognitively oriented work on the nature of mental representations.
- One of his students, John Anderson, worked with Bower on a very influential semantic network model (HAM), that was described in their 1973 book, Human Associative Memory. Later Anderson combined it with a production system component in ACT and its predecessors
    - ACT

Neuroscience

Research into the brain was long thought to be relevant to understanding mental processes. One line of research focused on deficits stemming from brain lesions, such as Broca's classic nineteenth century work lining articulate speech to what is now called Broca's area.

- Communication Problems
- Frontal lobe
- Language
- Organization of the Brain

One of the fruitful products of this engagement in the 1940s to 1960s was the development of neural networks, a kind of computational modeling pioneered by neurophysiologist Warren McCulloch and logician Walter Pitts.

Donald Hebb proposed to build cell-assemblies by strengthening connections between neurons that fired simultaneously, a technique still in use

Oliver Selfridge had layers of units competing in parallel to recognize patterns in his Pandemonium simulation

Frank Rosenblatt built layered networks that learned through error correction (Perceptrons).

Neural networks lost influence due to a devastating critique of Perceptrons by Minsky and Seymour Papert in 1969, but were revived when more effective techniques became available in the 'new connectionism' of the 1980s and beyond.

Linguistics

Structuralist linguists such as Franz Boas, Edward Sapir, and the positivist Leonard Bloomfield focused on lower-level structural units (phonemes and morphemes).

In the 1950s, post-Bloomfieldian Zellig Harris turned his attention to syntax and introduced the idea of transformations that normalized complex sentences by relating them to simpler kernel sentences.

- This idea launched a revolution in linguistics when it was further developed by Harris's student Noam Chomsky
  - Controversies in cognition and communication The History of Cognitive Psychology
- In his 1957 Syntactic Structures, Chomsky proposed the idea of a grammar as a generative system—a set of rules that would generate all and only members of the infinite set of grammatically well-formed sentences of a human language—and argued that finite state and phrase structure grammars, though generative, were inadequate.
- A series of transformations was needed to obtain an appropriate surface structure from an initial deep structure created by means of phrase structure rules.

Kuhn - The Structure of Scientific Revolutions

> The book introduces the concept of scientific revolutions and challenges the traditional view of scientific progress as a cumulative and linear process.

Paradigm Shifts:

- Kuhn argues that scientific development is marked by periods of normal science, where scientists work within a shared framework or paradigm. However, occasionally, anomalies or problems arise that cannot be solved within the existing paradigm. When these anomalies accumulate, it may lead to a crisis in the scientific community.

Scientific Revolutions:

- The resolution of a crisis often results in a paradigm shift or scientific revolution. During a revolution, there is a fundamental change in the underlying assumptions, theories, and methodologies that govern scientific inquiry. This shift is not a gradual accumulation of knowledge but rather a rapid and radical transformation.

Incommensurability:

- Kuhn introduces the concept of incommensurability, which means that paradigms are so different that they are not directly comparable. The terms and concepts used within one paradigm may not have the same meaning or reference in another, making it challenging for scientists from different paradigms to communicate effectively.

Normal Science:

- Between revolutions, scientists engage in what Kuhn calls normal science, where they work within the established paradigm, solving puzzles and conducting research guided by the accepted theories. Normal science aims to further refine and extend the existing paradigm.

Social and Psychological Factors:

- Kuhn emphasizes the role of social and psychological factors in scientific development. The scientific community plays a crucial role in determining which theories are accepted or rejected. Scientists are often deeply invested in the prevailing paradigm, and resistance to change can be strong.

Section I

"An apparently arbitrary element, compounded of personal and historical accident, is always a formative ingredient of the beliefs espoused by a given scientific community at a given time"

From Stanford Encyclopedia

The Development of Science

Thomas Kuhn - The Development of Science

The opinions before Kuhn was that, science develops by the addition of new truths to the stock of old truths, or the increasing approximation of theories to the truth, and in the odd case, the correction of past errors. Such progress might accelerate in the hands of a particularly great scientist, but progress itself is guaranteed by the scientific method.

According to Kuhn the development of a science is not uniform but has alternating '**normal**' and '**revolutionary**' (or '**extraordinary**') phases.

- The revolutionary phases are not merely periods of accelerated progress, but differ qualitatively from normal science.

Normal science does resemble the standard cumulative picture of scientific progress, on the surface at least. Kuhn describes normal science as 'puzzle-solving' (1962/1970a, 35–42).
While this term suggests that normal science is not dramatic, its main purpose is to convey the idea that like someone doing a crossword puzzle or a chess problem or a jigsaw, the puzzle-solver expects to have a reasonable chance of solving the puzzle, that his doing so will depend mainly on his own ability, and that the puzzle itself and its methods of solution will have a high degree of familiarity.

- A puzzle-solver is not entering completely uncharted territory. Because its puzzles and their solutions are familiar and relatively straightforward, normal science can expect to accumulate a growing stock of puzzle-solutions.

Revolutionary science, however, is not cumulative in that, according to Kuhn, scientific revolutions involve a revision to existing scientific belief or practice (1962/1970a, 92). Not all the achievements of the preceding period of normal science are preserved in a revolution, and indeed a later period of science may find itself without an explanation for a phenomenon that in an earlier period was held to be successfully explained. This feature of scientific revolutions has become known as 'Kuhn-loss' (1962/1970a, 99–100).

If, as in the standard picture, scientific revolutions are like normal science but better, then revolutionary science will at all times be regarded as something positive, to be sought, promoted, and welcomed. Revolutions are to be sought on Popper's view also, but not because they add to positive knowledge of the truth of theories but because they add to the negative knowledge that the relevant theories are false. Kuhn rejected both the traditional and Popperian views in this regard.

He claims that normal science can succeed in making progress only if there is a strong commitment by the relevant scientific community to their shared theoretical beliefs, values, instruments and techniques, and even metaphysics. This constellation of shared commitments Kuhn at one point calls a 'disciplinary matrix' (1970a, 182) although elsewhere he often uses the term '**paradigm**'.

- Because commitment to the disciplinary matrix (paradigm, red) is a pre-requisite for successful normal science, an inculcation of that commitment is a key

element in scientific training and in the formation of the mind-set of a successful scientist.

- Basically Kuhn emphasizes a conservative attitude in science which distinguishes him from Popper and his depiction of a scientist forever trying to refute his own theories

- This conservative resistance to the attempted refutation of key theories means that revolutions are not sought except under extreme circumstances.

Popper's philosophy requires that a single reproducible, anomalous phenomenon be enough to result in the rejection of a theory (Popper 1959, 86–7).

- Kuhn's view is that during normal science scientists neither test nor seek to confirm the guiding theories of their disciplinary matrix (paradigm). Nor do they regard anomalous results as falsifying those theories. (It is only speculative puzzle-solutions that can be falsified in a Popperian fashion during normal science (1970b, 19).)

Rather, anomalies are ignored or explained away if at all possible. It is only the accumulation of particularly troublesome anomalies that poses a serious problem for the existing disciplinary matrix.

- A particularly troublesome anomaly is one that undermines the practice of normal science. For example, an anomaly might reveal inadequacies in some commonly used piece of equipment, perhaps by casting doubt on the underlying theory.
- If much of normal science relies upon this piece of equipment, normal science will find it difficult to continue with confidence until this anomaly is addressed. A widespread failure in such confidence Kuhn calls a '**crisis**' (1962/1970a, 66–76).

The most interesting response to **crisis** will be the search for a **revised disciplinary matrix**, a revision that will allow for the elimination of at least the most pressing anomalies and optimally the solution of many outstanding, unsolved puzzles. Such a revision will be a **scientific revolution.**

- According to Popper the revolutionary overthrow of a theory is one that is logically required by an anomaly.
- According to Kuhn however, there are no rules for deciding the significance of a puzzle and for weighing puzzles and their solutions against one another.
    - The decision to opt for a revision of a disciplinary matrix is not one that is rationally compelled; nor is the particular choice of revision rationally compelled.

Kuhn states that science does progress, even through revolutions (1962/1970a, 160ff). The phenomenon of Kuhn-loss does, in Kuhn's view, rule out the traditional cumulative picture of progress.

- The revolutionary search for a replacement paradigm is driven by the failure of the existing paradigm to solve certain important anomalies.
- Any replacement paradigm had better solve the majority of those puzzles, or it will not be worth adopting in place of the existing paradigm.
- At the same time, even if there is some Kuhn-loss, a worthy replacement must also retain much of the problem-solving power of its predecessor (1962/1970a, 169).

Rejecting a teleological view of science progressing towards the truth, Kuhn favors an evolutionary view of scientific progress (1962/1970a, 170–3)

- The evolutionary development of an organism might be seen as its response to a challenge set by its environment. But that does not imply that there is some ideal form of the organism that it is evolving towards.
    - Analogously, science improves by allowing its theories to evolve in response to puzzles and progress is measured by its success in solving those puzzles; it is not measured by its progress towards to an ideal true theory.

While evolution does not lead towards ideal organisms, it does lead to greater diversity of kinds of organism.
According to this account, the revolutionary new theory that succeeds in replacing another that is subject to crisis, may fail to satisfy all the needs of those working with the earlier theory.

- One response to this might be for the field to develop two theories, with domains restricted relative to the original theory (one might be the old theory or a version of it).
- This formation of new specialties will also bring with it new taxonomic structures and so leads to **incommensurability**.

Paradigms in Cognitive Science

- Progress of science is not linear
- Science is done within paradigms
- Paradigms shape what phenomena can be studied and how these phenomena can be interpreted
- Paradigms are often *incommensurable*

Learning Goals

- Understanding Kuhns idea of paradigms and paradigm shifts
- Understanding why paradigms may be *incommensurable*
- Capability to reflect on what amounts to truth

Paradigms

- Behaviorism
  - Precognitive
- Computationalism
  - Cognition is symbol manipulation and can be instantiated in any architecture
- Connectionism
  - Cognition is the activity of connected networks
- Predictive processing
  - Perception is causal inference (i.e. perception is cognition)

A shift in outlook from normative

Normative

- Logical positivism
  - Acceptability of a theory depends on its correspondence to evidence
- Falsification
  - Holding on to theories depends on them withstanding falsification

Strongly put: If you're not following these norms, you're not doing science

A simple model

Immature science --> Normal Science --> Crisis --> Revolution --> Resolution (back to normal science)

Immature science

- Accumulation of effects to be explained

- Lack of agreed upon standards
  - Kuhn argues that the social sciences are in this state
- Lack of standard theory means that observations will be understood differently
  - Important insight of Kuhn

Immature Science

- Accumulation of effects to be explained
  - e.g. psychology has a bunch of peculiar effects (McGurk effect, Stroop Effect, visual illusions)
- Lack of agreed upon standard theories
  - Kuhn argues that the social sciences are in this state
- Lack of a standard theory means that observations will be understood differently
  - Important insight of Kuhn

Two key insights

1. There is no such thing as neutral observation
2. A theory, T_1 , cannot be verified or falsified without considering the grander network of theories, $\{T_1, T_2, \ldots, T_n\}$ that T_1 is part of

There is no such thing as neutral observation
All observation is embedded in theory

- Relates to the point about auxiliary theories (Duhem-Quine thesis)

THE DRESS and Kuhn

The internet dress is perceived differently from person to person, because the dress is framed and lit perfectly, so that your brain can have one or two theories of the lighting context

- Your perception of the dress is global
  - If the GOLD interpretation wins then it's gold everywhere
  - Paradigms holistically structure how evidence is interpreted
- Your perception of the dress can change, but that is also global

    - If it starts to look blue, it will look blue everywhere
    - If your paradigm changes, then it will change for ALL the evidence
- The input into your sensory system can be highly consistent with both interpretations
    - The same evidence can imply more than one paradigm, and the evidence can even look different dependent on which paradigm you assume
- You can only have one interpretation at a time
    - Paradigms are *incommensurable*
- • The other interpretation can appear irrational
    - Relativism? Progress?

Normal Sciences

A ruling paradigm states the accepted theory of the day
Thus also determines the “correct” interpretation of observations

Essential characteristics of normal science (or paradigmatic science)

1. Their achievement was sufficiently unprecedented to attract an enduring group of adherents away from competing modes of scientific activity
    1. Cognition was seen as taboo (George Mandler in Baars 1986)
2. it was sufficiently open-ended to leave all sorts of problems for the redefined group of practitioners to resolve

Note the shift away from logical considerations towards **historical** and **sociological** considerations

Quick run-down of lectures

Week 11 – *Behaviorism* (Skinner the big bad)

- The proper study of the mind is the study of behavior
- Behavior is operations on the environment
- To predict and control behavior, one needs to control the environment

First essential characteristic

- Their achievement was sufficiently unprecedented to attract an enduring group of adherents away from competing modes of scientific activity

- 1940s–1960s: Behaviorism dominates at least North American psychology

Cognition = taboo

> ‘... cognition was a dirty word for us [...] because cognitive psychologists were seen as fuzzy, hand-waving, imprecise people who never really did anything that was testable’ (George Mandler quoted in Baars 1986, my italicization).

Second essential characteristic

- [...] it was sufficiently open-ended to leave all sorts of problems for the redefined group of practitioners to resolve
  - How to most efficiently speed up learning?
  - How much can the reward be delayed in reinforcement learning?
  - Is reinforcement or punishment most effective for learning?
- Kuhn calls this *puzzle solving*

Before Kuhn

We take a step back and take another look at how things were before Kuhn

Key distinction: (kennen/wissen)

- Knowledge by acquaintance
  - Things we know by a causal connection to the world
  - Things we know by experience
  - Non-inferential knowledge
- Knowledge by description
  - Things we don’t know by experience
  - Inferential knowledge

Meaning Holism

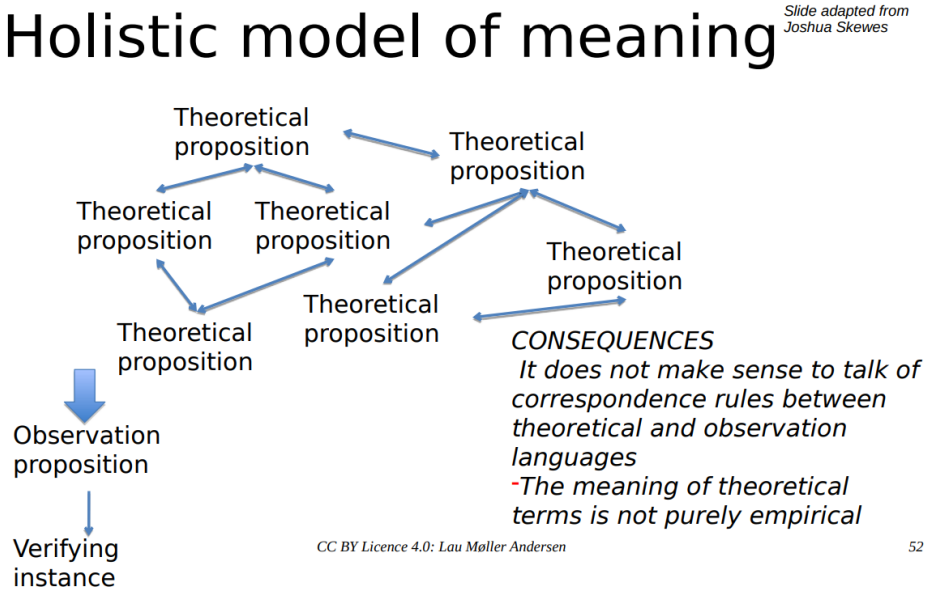
Kuhn disagrees with previous scientific reasoning
He argues for what's called **meaning holism***

> “…our statements about the external world face the tribunal of sense experience not individually but only as a corporate body…”

- The meaning of a term in a theory *is* dependent on its functional role in that theory
- We can’t verify theoretical propositions, we can only verify whole theories
  - My own words: We verify the whole (holism) i.e. the whole theory not individual propositions

Famous paper: “Two Dogmas of Empiricism”

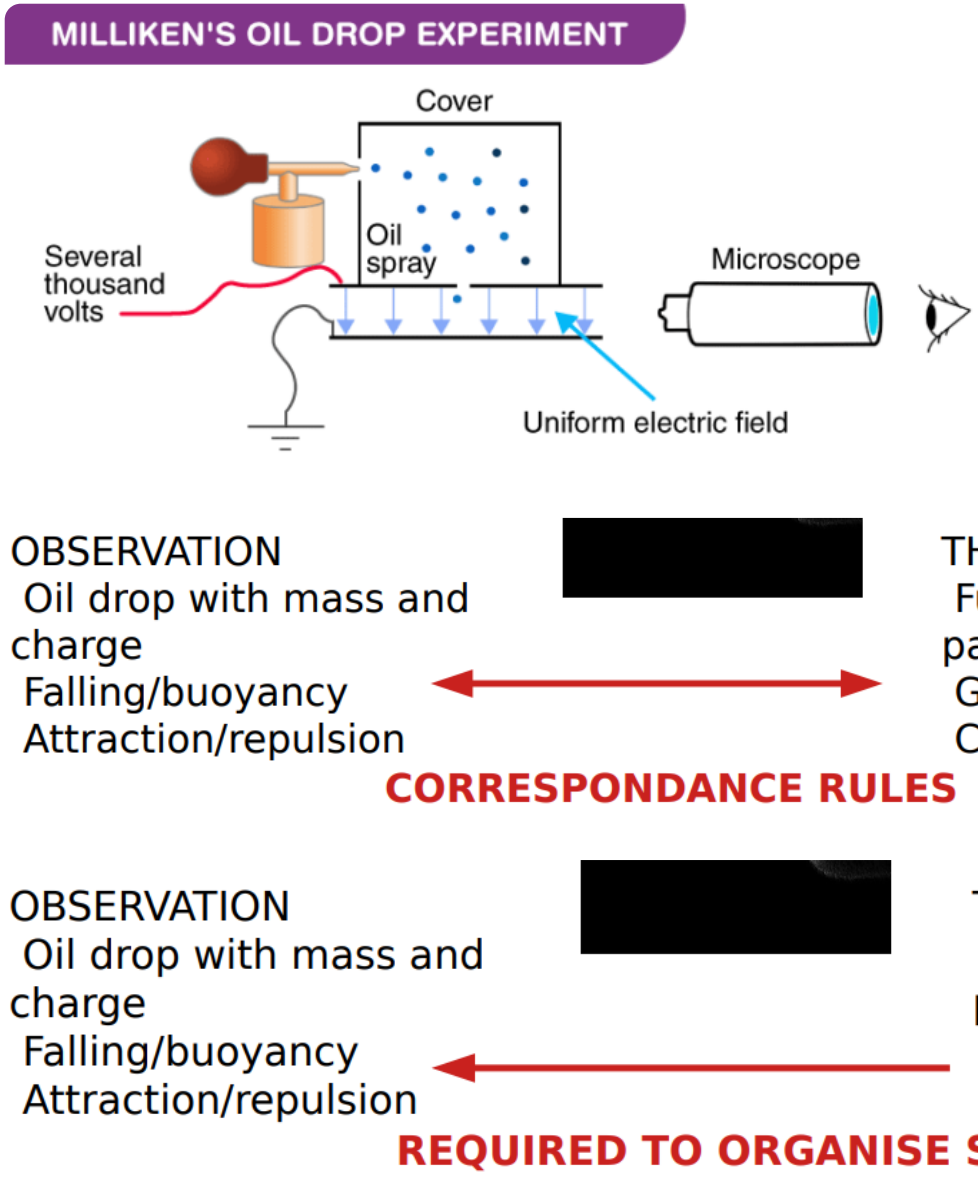
Holistic model of meaning



CONSEQUENCES:

- It does not make sense to talk of correspondence rules between theoretical and observation languages
  - The meaning of theoretical terms is purely empirical

Example: Millikan oil drop experiment



A consequence of holism is incommensurability

Crisis

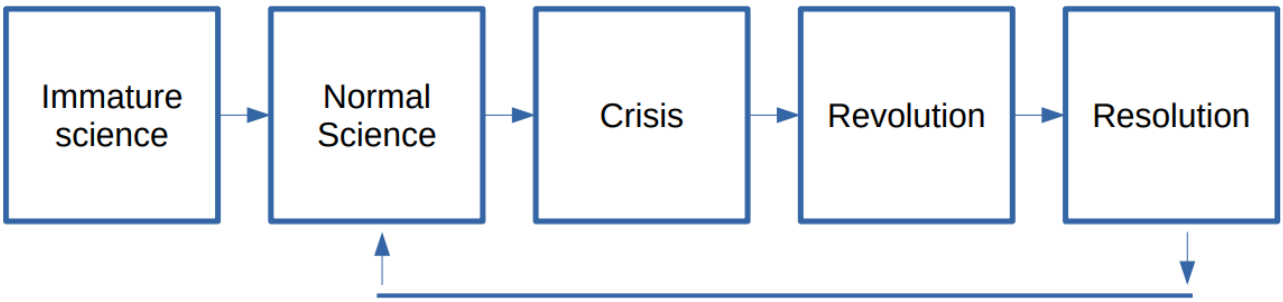
Examples that can bring it about

- We are running out of puzzles to solve, i.e. paradigm not fruitful in generating new hypotheses
- Our puzzle solving strategy/paradigm doesn't work on many of the interesting problems anymore
- Technological advances foster new ways of thinking
- Social climate might change the questions that are being asked

Resolution

> "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." (quoted in Kuhn, 1973, p.151)

A simple model (again)



B. F. Skinner - Behaviorism

> "In these excerpts from Science and Human Behavior, Skinner presents a comprehensive account of the theory of behaviorism and argues for the superiority of his approach to various other ways of understanding "mental phenomena"

Skinner claims it is unscientific and fruitless to try to explain human behavior by reference to inner causes, instead we should seek the causes outside of the individual in the environment

Skinner maintains that all human behavior can be explained by reference to three different stimulus-response relations:

- **Unconditioned reflexes**
  - A narrow range of human behavior such as tearing in the presence of onions
  - Automatic/reflexes
- **Conditioned reflexes**
  - Somewhat wider class, e.g. Pavlov's dogs salivating by the ring of a bell after it has been paired with the unconditioned stimulus of food
- **Operant conditioning**
  - The vast majority of human behavior; when a particular behavior is rewarded or "reinforced"

Inner Causes

Neural causes: Skinner doesn't ascribe the nervous system much responsibility for behavior

- "We may note here that we do not have and may never have this sort of neurological information which is needed to predict a specific instance of behavior"
- "The causes to be sought in the nervous system are therefore of limited usefulness in the prediction and control of specific behavior"

Psychic inner causes: Skinner discredits the act of explaining behavior in terms of an inner agent "mental"/"psychic"

- Disregards Freudian concepts of ego, superego, id and the unconscious
- "Direct observation of the mind comparable with the observation of the nervous system has not proved feasible"
- "Introspective psychology no longer pretends to supply direct information about events which are the causal antecedents, rather than the mere accompaniments of behavior. It defines its "subjective" events in ways which strip them of any usefulness in causal analysis"

The Variables of which Behavior is a Function

Skinner creates an example with a person drinking water

- Gives reasons for why a subject may or may not drink the water
- "In both cases, either for accurate prediction or control, we must investigate the effect of each variable quantitatively with the methods and techniques of a laboratory science"
- "We must know the history of our subject with respect to the behavior of drinking water and if we cannot eliminate social factors from the situation, then we must know the history of his personal relations to people resembling the experimenter"

To what extent is it helpful to be told: "he drinks because he is thirsty"?

- If to be thirsty means nothing more than to have a tendency to drink, this is mere redundancy
- If it means that he drinks because of a state of thirst, an inner causal event is invoked. If this state is purely inferential-- if no dimensions are assigned to it

which would make direct observation possible-- it cannot serve as an explanation
- But if it has physiological or psychic properties, what role can it play in a science of behavior?

1. An operation performed upon the organism from without-- for example, water deprivation
2. An inner condition-- physiological or psychic thirst
3. A kind of behavior, drinking

The second variable would be preferred as it is non-historic (the first may lie in the past, the second is current)
Sometimes we infer the second from the third, an animal is judged to be thirsty if it drinks

- In this case the explanation is spurious

The second link is useless in the *control* of behavior unless we can manipulate it
We usually set up the second link through the first; we make an animal thirsty

The most objectionable practice is to follow the causal sequence back only as far as a hypothetical second link. It is no help to be told that to get an organism to drink we are simply to "make it thirsty" unless we are also told how it is to be done

The objection to inner states is not that they do not exist, but that they are not relevant in a functional analysis

A Functional Analysis

We undertake to predict and control the behavior of the individual organism (the dependent variable)
Our independent variable is the cause of behavior (external conditions of which behavior is a function)

"We shall see later that although such terms as "meaning" and "intent" appear to refer to properties of behavior, they usually conceal references to independent variables. This is also true of "aggressive" "friendly" etc. and other terms which appear to describe properties of behavior but in reality refer to its controlling relations."

Reflex Action

The external agent came to be called a *stimulus*. The behavior controlled by it to be called a *response*. Together they comprised what was called a *reflex*

Many characteristics of the relation have been studied quantitatively.

- The time which elapses between stimulus and response (the *latency*)
- Magnitude of the response as a function of the intensity of the stimulus
- Other conditions have been found to be important in completing the account, e.g. repeated rapid elicitation of the action may cause a reflex to be *fatigued*

The Range of Reflex Action

If we were to assemble all the behavior which falls into the pattern of the simple reflex, we should have only a very small fraction of the total behavior of the organism. It is neither plausible nor expedient to conceive of the organism as a complicated jack-in-the-box with a long list of tricks, each of which may be evoked by pressing the proper button.

- The greater part of the behavior of the intact organism is not under this primitive sort of stimulus control.
- The environment affects the organism in many ways which are not conveniently classed as "stimuli" and even in the field of stimulation only a small part of the forces acting upon the organism elicit responses in the invariable manner of reflex action

To ignore the principle of the reflex entirely, however, would be equally unwarranted

Conditioned Reflexes

On Pavlov:
"Without influencing the dog in any other way, he could sound a tone and insert food into the dog's mouth. In this way he was able to show that the tone *acquired* its ability to elicit secretion... He has replaced the "psyche" of psychic secretion with certain objective facts in the recent history of the organism"

The process of conditioning as Pavlov reported in his book *Conditioned Reflexes* is a process of *stimulus substitution*

- A previously neutral stimulus acquires the power to elicit a response which was originally elicited by another stimulus

- The change occurs when the neutral stimulus is followed or "reinforced" by the effective stimulus

Pavlov studied quite a few variables among which the converse process, where the conditioned stimulus loses its power to evoke the response when it is no longer reinforced, a process which he called *extinction*

Learning Curves

Skinner talks about E. L. Thorndike and his cat experiment here (see Thorndike)

- Thorndike found that when a cat was put into such a box again and again, the behavior which led to escape tended to occur sooner and sooner until eventually escape was as simple and quick as possible
    - The cat had solved its problem as well as if it were a "reasoning" human being ... yet Thorndike observed no "thought-process"

By noting the successive delays in getting out of the box and plotting them on a graph, he constructed a "learning curve"

- This early attempt to show a quantitative process in behavior, similar to the processes of physics and biology, was heralded as an important advance

Operant Conditioning

To get at the core of Thorndike's **Law of Effect** we need to clarify the notion of "probability of response"

- Skinner writes a lot on what constitutes a habit or predisposition

In an example of conditioning a bird to raise its head above a certain line in order to receive food Skinner writes this:

- "The statement that the bird "learns that it will get food by stretching its neck" is an inaccurate report of what has happened"
    - The barest possible statement of the process is this: we make a given consequence contingent upon certain physical properties of behavior (the upward movement of the head) and the behavior is then observed to increase in frequency

A response which has already occurred cannot be predicted or controlled. We can only predict that *similar* responses will occur in the future. The unit of a predictive science is there not a response but a class of responses. The word "*operant*" will be used to describe this class. The term emphasizes the fact that the behavior *operates* upon the environment to generate consequences

Pavlov himself called all events which strengthened behavior "reinforcement" and all the resulting changes "conditioning". In the Pavlovian experiment, however a reinforcer is paired with a *stimulus* whereas in operant behavior it is contingent upon a response

- Operant reinforcement is therefore a separate process

> "The change in frequency with which the head is lifted to this height is the process of *operant conditioning*"

Reading Questions

Skinner:

1. What is the main difference between Classical Conditioning and Operant Conditioning?
2. What does it mean that behaviorism is *externalist*? What is its stance on inner causes?
3. If we find inner causes, does it matter for *predicting* and *controlling* an organism's behavior?

Answers:

1. the main difference between classical and operant conditioning lies in the types of associations being made: classical conditioning involves associating stimuli with involuntary responses, while operant conditioning involves associating behaviors with their consequences (reward/punishment).
2. It attributes all behavior to external factors (environmental influences). Skinner's behaviorism completely disregards inner causes, deeming them to be irrelevant in functional analysis
3. Not according to Skinner no

Behaviorism Lecture

Behaviourism

- The proper study of the mind is the study of behaviour
- Behaviour is operations on the environment
- To explain, predict and control behaviour, one needs to control the environment

Theme

- How behaviorism provided the (proto)-paradigm for cognitive science, and why it eventually was abolished

Learning goals

1. Understanding how behaviourism is an externalist paradigm
2. Understanding what operant conditioning is, and how it differs from classical conditioning
3. Understanding the connection to empiricism and logical positivism
4. Being able to critically reflect on the assumptions of behaviourism
5. Being able to see how behaviourism is still relevant today

Essential characteristics of paradigms

1. Their achievement was sufficiently unprecedented to attract an enduring group of adherents away from competing modes of scientific activity 1940s–1960s: Behaviourism dominates at least North American psychology
2. it was sufficiently open-ended to leave all sorts of problems for the redefined group of practitioners to resolve (How to most efficiently speed up learning, reinforcement/punishment most effective for learning) *Kuhn's puzzle solving*

Features of behaviorism

- Empiricist
  - Emphasis on observation
- Positivist
  - Observations need to be intersubjective and external
  - Independent and dependent variables instead of cause and effect
- Instrumentalist
  - Focus on predicting and controlling

- No intermediaries necessary between environment and behavior

Empiricism, positivism and causation

- A treatise of human nature:
  - Cause is unobservable, so does not belong to an empiricist framework
- Science and Human Behavior:
  - Replace cause and effect with a functional analysis of behaviour: predicting and controlling behaviour through independent and dependent variables

Functional analysis:

- Relations between variables
  - Independent variable replaces cause
  - Dependent variable replaces effect
  - Independent variable ⇒ (intermediary variable) ⇒ dependent variable
- Doing science
  - Not understanding causes
  - Using knowledge of independent variables to understand dependent variables
  - What is understanding?
    - Prediction
    - Control

Variables according to behaviorism:
In a broad sense:

```
independent  ⇒ intermediary ⇒ dependent

environment  ⇒ inner causes ⇒ behaviour
```

more narrowly

```
  schedules
     of        ⇒ inner causes ⇒   frequency
                                      of
reinforcement                    behaviour
```

Not quite sure how the inner causes fit into this? Thought behaviorism excluded all considerations of inner causes?

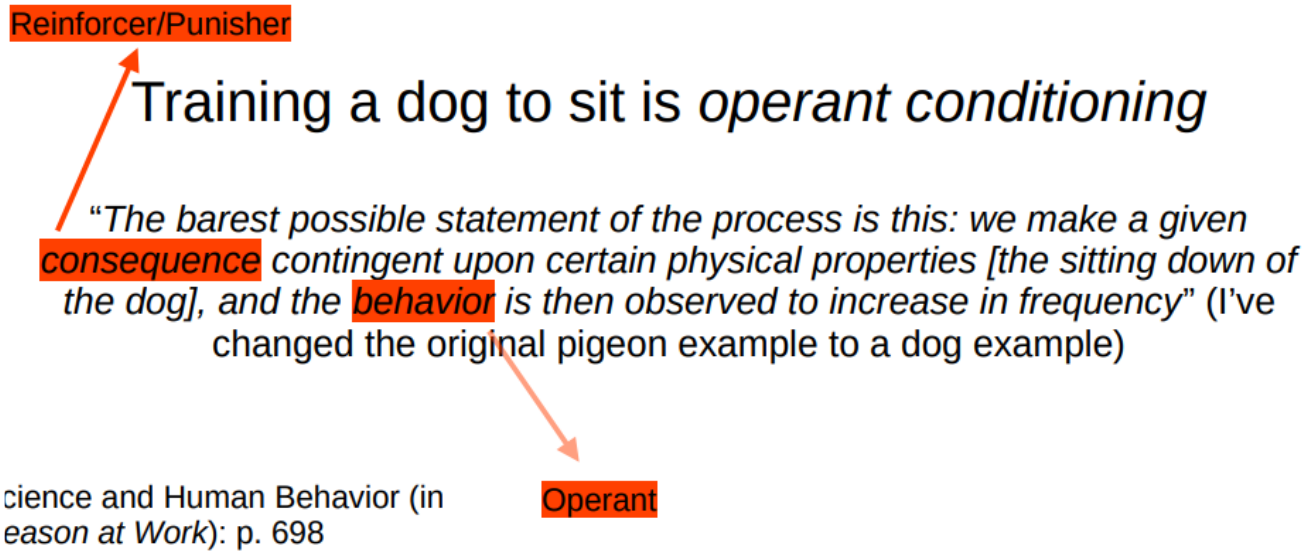
How to operationalise reinforcement (and punishment)

- *Reinforcer*: something that increases the frequency of a behavior
- *Punisher*: something that decreases the frequency of a behavior

Operationalization: describing a concept in terms of (in)dependent variables

Some terminology:

- *Response*: single instance of (not) sitting down
- *Operant*: the behaviour of (not) sitting down generally



Doing psychology (remember): according to Carnap

#Philosophy

Introduction. Physical Language and Protocol Language

Opens quite strongly:

> "IN WHAT FOLLOWS, we intend to explain and to establish the thesis that every sentence of psychology may be formulated in Physical language. To express this in the material mode of speech: all sentences of psychology describe physical occurrences, namely, the physical behavior of humans and other animals."

and

> "This is a sub-thesis of the general thesis of *physicalism* to the effect that physical language is a *universal language*, that is, a language into which every sentence may be translated"

If our thesis is correct, the generalized sentences of psychology, the laws of psychology, are also translatable into physical language. They are thus physical laws.

The Forms of Psychological Sentences

- A singular psychological sentence e.g. “Mr. A was angry at noon yesterday” (an analogue of the physical sentence, “Yesterday at noon the temperature of the air in Vienna was 28 degrees centigrade”), is concerned with a particular person at a particular time
- General psychological sentences have various forms, of which the following two are perhaps the most important
    1. A sentence may describe a specific quality of a specific kind of event, e.g. “An experience of surprise always (or: always for Mr. A, or: always for people of such and such a society) has such and such a structure.” A physical analogy would be: “Chalk (or: chalk of such and such a sort) always is white.”
    2. universal-conditional statements concerning sequences of events, that is, of causal laws. For instance,
        1. “When, under such and such circumstances, images of such and such a sort occur to a person (or: to Mr. A, or: to anyone of such and such a society), an emotion of such and such a sort always (or: frequently, or: sometimes) is aroused.” A physical analogy would be: “When a solid body is heated, it usually expands.”

Phenomenology claims to be able to establish universal synthetic sentences which have not been obtained through induction. These sentences about psychological qualities are, allegedly, known either *a priori* or on the basis of some single illustrative case

- In our view, knowledge cannot be gained by such means. We need not, however, enter upon a discussion of this issue here, since even on the view of

phenomenology itself, these sentences do not belong to the domain of psychology.

In physics it sometimes seems to be the case that a general law is established on the basis of some single event. For instance,

- if a physicist can determine a certain physical constant, say, the heat-conductivity of a sample of some pure metal, in a single experiment, he will be convinced that, on other occasions, not only the sample examined but any similar sample of the same substance will, very probably, be characterizable by the same constant.
- But here too induction is applied. As a result of many previous observations the physicist is in possession of a universal sentence of a higher order which enables him in this case to follow an abbreviated method. This higher-order sentence reads roughly: “All (or: the following) physical constants of metals vary only slightly in time and from sample to sample.”

The situation is analogous for certain conclusions drawn in psychology. If a psychologist has, as a result of some single experiment, determined that the simultaneous sounding of two specific notes is experienced as a dissonance by some specific person A, he infers (under favorable circumstances) the truth of the general sentence which states that the same experiment with A will, at other times, have the same result

- Here too the inference from a singular sentence to a general one is only apparent. Actually, a sentence inductively obtained from many observations is brought into service here

It thus remains the case that every general sentence is inductively established on the basis of a number of singular ones.

Finally, we must consider sentences about psycho-physical interrelations, such as for instance, the connection between physical stimulus and perception. These are likewise arrived at through induction, in this case through induction in part from physical and in part from psychological singular sentences. **The most important sentences of gestalt psychology belong also to this kind.**

General sentences have the character of hypotheses in relation to concrete sentences, that is, the testing of a general sentence consists in testing the concrete sentences which are deducible from it. A general sentence has content insofar and only insofar as the concrete sentences deducible from it have content. Logical analysis must therefore primarily be directed towards the examination of the latter sort of sentences

Sentences About Other Minds

The epistemological character of a singular sentence about other minds will now be clarified by means of an analogy with a sentence about a physical property, defined as a disposition to behave (or respond) in a specific manner under specific circumstances (or stimuli).

A Sentence about a property of a physical substance

Example: I assert the sentence P_1: “This wooden support is very firm.”

A Sentence about a condition of some other mind.

Example: I assert the sentence P_1: "Mr. A is now excited.”

These are two different ways in which sentence P_1 may be derived. We shall designate them as the “ rational ” and the “ intuitive ” methods

The rational method consists of inferring P_1 from some protocol sentence p_1 (or from several like it), more specifically, from a perception-sentence

- about the shape and color of the wooden support.
- about the behavior of Mr. A, e.g. about his facial expressions, his gestures, etc., or about physical effects of A’s behavior, e.g. about characteristics of his handwriting

In order to justify the conclusion, a major premise O is still required, namely the general sentence which asserts that

- when I perceive a wooden support to be of this color and form, it (usually) turns out to be firm. (A sentence about the perceptual signs of firmness.)
- when I perceive a person to have this facial expression and handwriting he (usually) turns out to be excited. (A sentence about the expressional or graphological signs of excitement.)

Lau says:

- $P_1 \Leftrightarrow (p_1 \wedge O)$
- "P1 is true if and only if both p1 and O are true simultaneously." It's expressing an equivalence relationship between P1 and the conjunction of p1 and O.
    - Creds to ChatGPT

The content of P_1 does not coincide with that of p_1, but goes beyond it. This is evident from the fact that to infer P_1 from p_1 O is required. The cited relationship between P_1 and p_1 may also be seen in the fact that under certain circumstances,

the inference from p_1 to P_1 may go astray. It may happen that, though p_1 occurs in a protocol, I am obliged, on the grounds of further protocols, to retract the established system sentence P_1. I would then say something like, “I made a mistake. The test has shown:

- that the support was not firm, even though it had such and such a form and color.”
- that A was not excited, even though his face had such and such an expression.”

In practical matters the intuitive method is applied more frequently than this rational one, which presupposes theoretical knowledge and requires reflection. In accordance with the intuitive method, P_1 is obtained without the mediation of any other sentence from the identically sounding protocol sentence p_2.

- “The support is firm.”
- “A is excited.”

Consequently, one speaks in this case of *immediate perceptions*

- of properties of substances, e.g., of the firmness of supports.
- of other minds, e.g., of the excitement of A

Here too we can best clarify the difference by considering the possibility of error. It may happen that, though p_2 occurs in my protocol, I am obliged, on the basis of further protocols, to retract the established system sentence P_1. I would then say “I made a mistake. Further tests have shown:

- that the support was not firm, although I had the intuitive impression that it was.”
- that A was not excited, although I had the intuitive impression that he was.”

Bunch of text wont write, so picture:

respective fields, the majority of them nowadays would give us thoroughly non-analogous answers. The identity of the content of P_2

and of the content of the physical sentence P_1 would be agreed to as a matter of course by all physicists.

and of the content of the psychological sentence P_1 would be denied by almost all psychologists (the exceptions being the radical behaviorists).

The contrary view which is most frequently advocated by psychologists is that, "A sentence of the form of P_1 asserts the existence of a state of affairs not identical with the corresponding physical structure, but rather, only accompanied by it, or expressed by it. In our example:

P_1 states that the support not only has the physical structure described by P_2, but that, besides, there exists in it a certain force, namely its *firmness*.

This firmness is not identical with the physical structure, but stands in some parallel relation to it in such a manner that the firmness exists when and only when a physical structure of the characterized sort exists.

Because of this parallelism one may consider the described reaction to certain stimuli—which is causally dependent upon that structure—to be an *expression* of firmness.
Firmness is thus an occult property, an obscure power which stands behind physical structure, appears in it, but itself remains unknowable."

P_1 states that Mr. A not only has a body whose physical structure (at the time in question) is described by P_2, but that—since he is a *psychophysical being*—he has, besides, a consciousness, a certain power or entity, in which that excitement is to be found.
This excitement cannot, consequently, be identical with the cited structure of the body, but stands in some parallel relation (or in some relation of interaction) to it in such a manner that the excitement exists when and only when (or at least, frequently when) a physical, bodily structure of the characterized sort exists.
Because of this parallelism one may consider the described reaction to certain stimuli to be an *expression* of excitement.

Excitement, or the consciousness of which it is an attribute, is thus an occult property, an obscure power which stands behind physical structure, appears in it, but itself remains unknowable."

P_2: a dispositional statement (logical behaviorism)

This view falls into the error of a hypostatization[1] as a result of which a remarkable duplication occurs: besides or behind a state of affairs whose existence is empirically determinable, another, parallel entity is assumed, whose existence is not determinable. (Note that we are here concerned with a sentence about other minds.) But—one may now object—is there not really at least one possibility of testing this claim, namely, by means of the protocol sentence p2 about the intuitive impression of the firmness of the support?

The objector will point out that this sentence, after all, occurs in the protocol along with the perception sentence p1. May not then a system sentence whose content goes beyond that of P2 be founded on p2? This may be answered as follows.

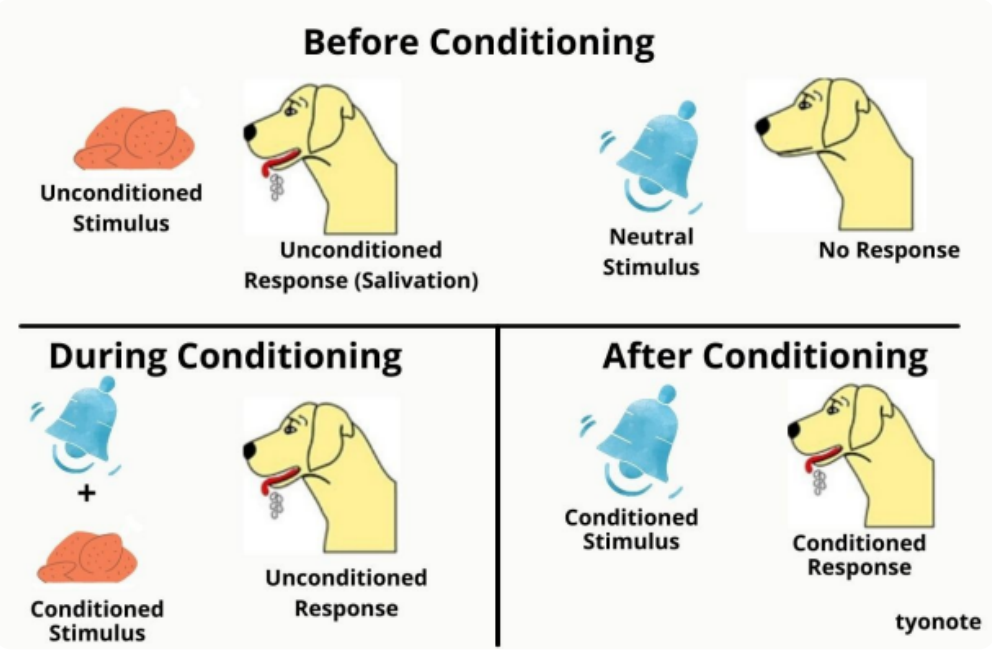
- A sentence says no more than what is testable about it. If, now, the testing of Pl consisted in the deduction of the protocol sentence p2, these two sentences would have the same content. But we have already seen that this is impossible

> There is no other possibility of testing PI except by means of protocol sentences like p1 of p2. If, now, the content of PI goes beyond that of P2, the component not shared by the two sentences is not testable, and is therefore meaningless. If one rejects the interpretation of PI in terms of P2, PI becomes a metaphysical pseudo-sentence.

1. Wtf↩

But what about Pavlov and his dogs? Isn't that behaviorism?

Conditioned reflexes (**classical conditioning**)



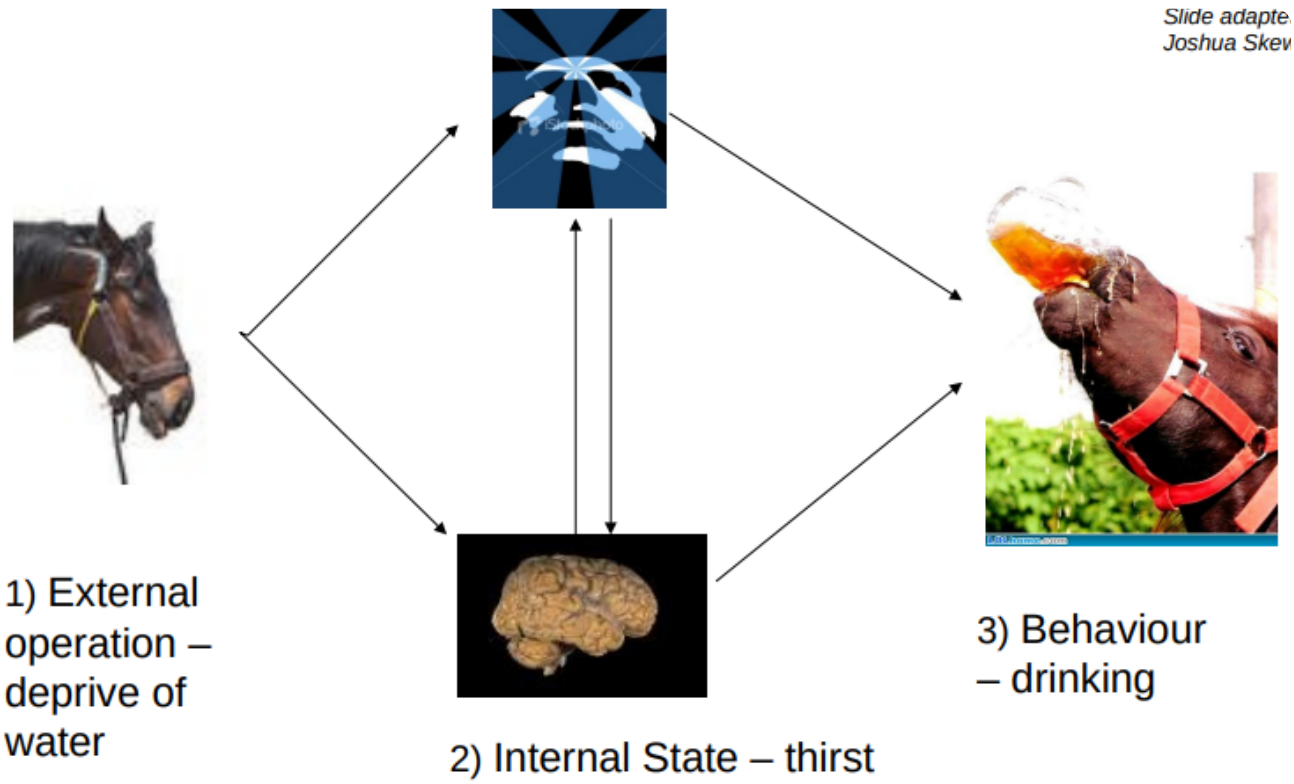
IMPORTANTLY
Classical conditioning needs a reflex to work on

WHEREAS
Operant conditioning is much more versatile

Remember the pigeons from last time?

- At face-value, the pigeons seem to read and play ping-pong the way that we do.
  - However, what do we need to know according to Skinner to explain, predict and control their sophisticated behavior?
  - WE NEED TO KNOW The pigeons' *histories* of operations

What about the inner causes? (intermediary variables)



Two arguments against intermediary variables

1. It is rarely possible to manipulate directly, e.g. “physiological or psychic thirst”
    - We can predict behaviour without it, i.e. by looking at the history of operations
    - Thus: intermediary variables are not necessary for predicting and controlling behaviour
2. Suppose intermediary variables were easy to manipulate directly
    - That manipulation would have to be included in an explanation
    - That manipulation is just another external operation
    - Explaining in terms of intermediary variables would be explanatorily empty without explaining this external operation
    - **Thus**: intermediary variables are not sufficient for predicting and controlling behaviour

Against inner causes:

- **Premise 1**: Intermediary variables are not necessary to scientific psychology
- **Premise 2**: Intermediary variables are not sufficient for scientific psychology
- **Conclusion**: Intermediary variables have no place in scientific psychology

How to do psychology

1. Observe a behaviour
2. Explain in terms of the history of operations
3. Perform further operations and predict outcomes
4. Perform those operations that lead to desired behaviour
5. Create a utopia with a technology of behaviour

What is an operation?

- Reinforcement and punishment

What is a history of operations?

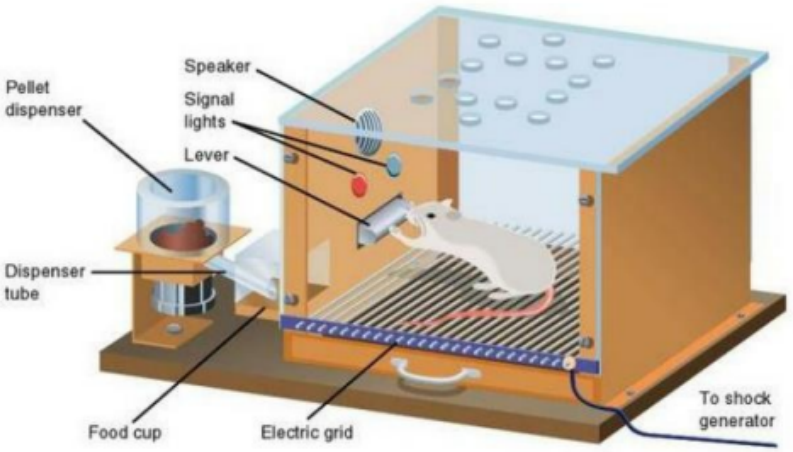
- Reinforcement schedule

What is a technology of behaviour

- A global Skinner box

Operant conditioning chamber

(Skinner box)



Skinner Skinned

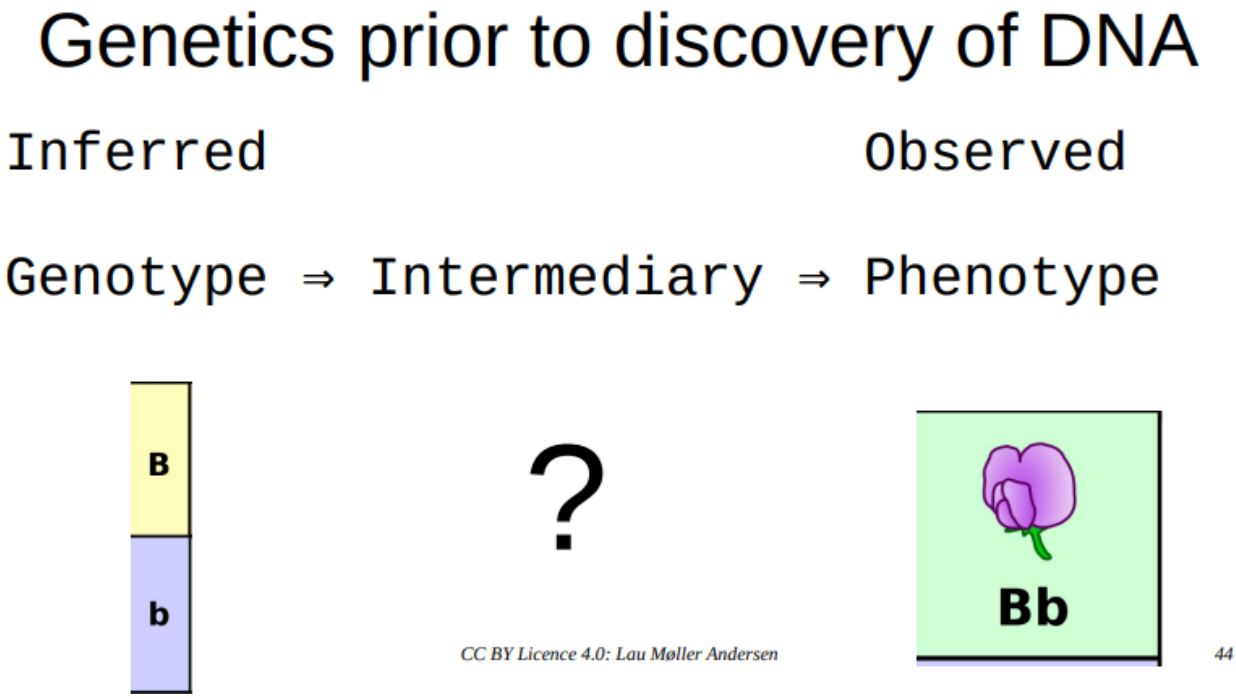
Revisiting the argument (adding 3rd premise)

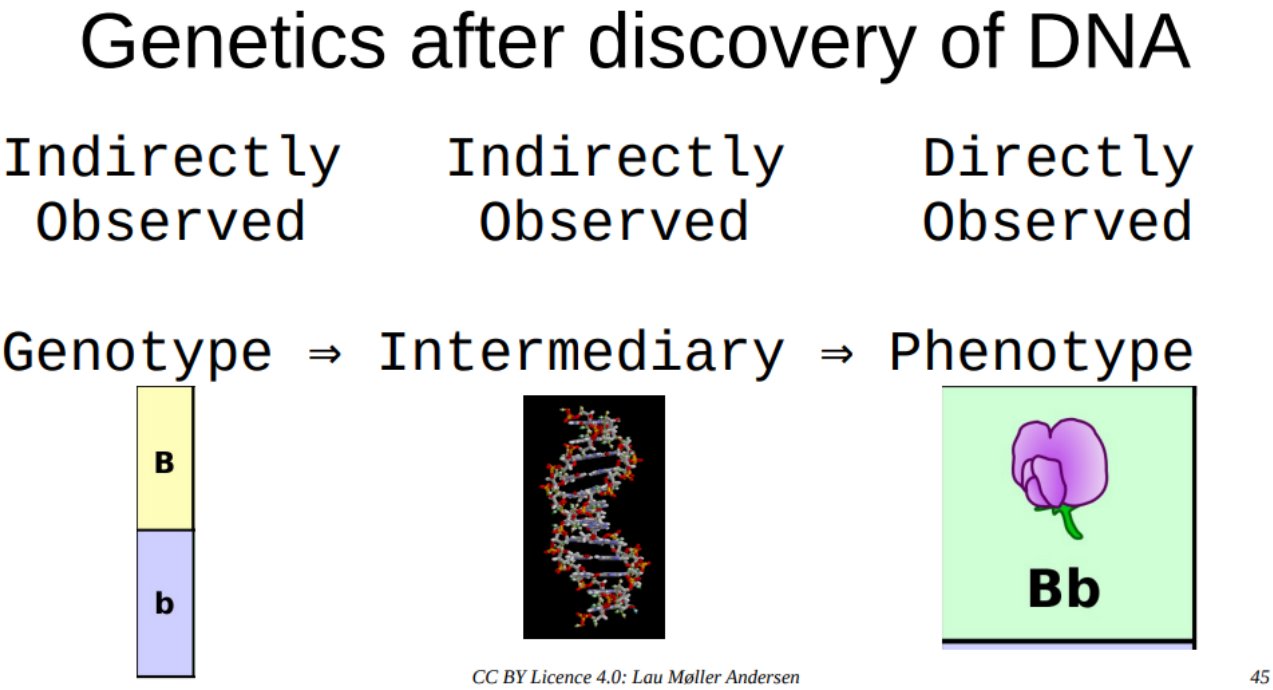
- Premise 3: Variables must be either necessary or sufficient to have a place in psychology

If you generalize this argument things become bad

- Premise 1: Chemical variables are not necessary to doing genetics
- Premise 2: Chemical variables are not sufficient for genetics
- Premise 3: Variables must be either necessary or sufficient to have a place in genetics
- Conclusion: Chemical variables have no place in genetics

Is this so?





Which intermediary variables (inner causes) are not relevant?

- Examples from Skinner:
  - *Vis viva*: “explains” the motion of a rolling stone
  - *Essence*: “explains” the chemical properties of bodies
  - *Vis medicatrix*: “explains” the healing of wounds
- From Molière via Dennett:
  - *Virtus dormitiva*: “explains” why opium make people sleepy

These are all inferential and question-begging

Question-begging: supporting a claim (a conclusion) with a premise that restates or presupposes the claim

Revisiting the argument (again?)

- Premise 1: Question begging variables are not necessary to scientific psychology
- Premise 2: Question begging variables are not sufficient for scientific psychology
- Premise 3: Variables must be either necessary or sufficient to have a place in psychology
- Conclusion: Intermediary variables have no place in scientific psychology

Final argument

- Premise 0: Intermediary variables are necessarily question-begging in psychological theories
- Premise 1: Question begging variables are not necessary to scientific psychology –
- Premise 2: Question begging variables are not sufficient for scientific psychology –
- Premise 3: Variables must be either necessary or sufficient to have a place in psychology
- Conclusion: Intermediary variables have no place in scientific psychology

Premise 0 (expanded)

- All theories employing mental terminology require a homunculus
- The homunculus used can always have the property required by the theory
- You can always come up with a mental explanation of any behaviour
- **Thus**: Intermediary variables are necessarily question begging in psychological theories

Dennett's comparison:

- Behaviourist theories require postulating a reinforcement schedule
- The schedule can always have the property required by the behaviour
- You can always come up with an external explanation of any behaviour
- **Thus**: external variables are necessarily question-begging in any behaviourist theory

1. **Watching a chess computer play a human**

- After watching many games, the externalist (behaviourist) could say something like:
  - “there is a high probability that the computer will move the queen to king’s bishop-4 because when stimulated by similar board positions in the past [...],

the computer has been reinforced for making similar [...] moves”

- The internalist might say something like:
  - “Now if I were the computer, knowing what I know and wanting what I want, which move would I believe to have the best consequences”

2. **Two chess computers play each other**

- Training (history of operations)
  - They learn overnight by playing each other multiple times
  - We do not see this and no record is kept
- Observation (behaviour)
  - We see them play the next morning

Predictions:

- Externalist: must invoke a history of operations even though it is by definition lost
- Internalist: must invoke rationality and intentionality (being about something)

A brief point about the relevance of behaviorism today

- Most animal studies use behaviourist methodologies to “instruct” their animals, i.e. they need to learn a task
- Only when certain capabilities cannot be explained behaviourally, can we invoke mental explanations (in animals)
- Case in point: metacognition: (knowing what you know and that you know)

Metacognition

- *Public mechanisms*
  - Environmental cue associations
    - e.g. do the stimuli themselves reveal this is a hard trial?
  - Behavioural cue associations
    - e.g. can the animal observe its own vacillation?
  - Response competition
    - Does the discrimination response compete with the metacognitive response?
- *Private mechanism*
  - Introspection

Lau's personal take

- don’t make fanciful mental explanations, if behaviourist explanations will do!

Dennett - Skinner Skinned

> This article is Daniel Dennett's response to Skinner's account of the theory of Behaviorism, see B. F. Skinner - Behaviorism

Dennett's first task is to try to fathom the reasons behind Skinner's theory. In particular, Dennett focuses on the question of why Skinner rejects explanations of human behavior which appeal to mental processes

- Dennett calls these explanations, which refer to things like the agent's beliefs, desires, reasonings, or reflections, "*intentional explanations*" or sometimes "*mentalistic explanations*"
- According to Dennett, Skinner has a fairly reasonable objection to this type of explanation, this being that mentalistic explanations are too easy and they do not increase our understanding

For example, if a friend were to answer "why are you reading this book" and you were to answer "Because I want to", your friend might not feel much of an explanation has been provided

- Dennett still maintains though that a beginnings of an explanation has been given, because this answer does rule out some possibilities
- Dennett believes that Skinner would be right only if mentalistic explanations had to stop at this level - with wants, desires etc., then they would be almost useless
- Dennett locates Skinners error in the belief that mentalistic explanations must terminate at this superficial level
    - If we deepen the explanation (what a want is and where it comes from) then the mentalistic explanation can be viewed as the first step in a serious and illuminating theory

Basically:
Dennett argues that Skinner's behaviorism oversimplifies the complexity of human behavior by reducing it to stimulus-response patterns. He contends that Skinner's model fails to adequately address the internal mental processes that influence behavior, such as beliefs, desires, intentions, and consciousness.

Furthermore, Dennett suggests that Skinner's emphasis on external reinforcement as the primary determinant of behavior neglects the role of internal states and cognitive processes. He posits that human behavior cannot be fully understood without

considering the role of mental representations and the interpretation of stimuli by the individual.

The Actual Article

> "In particular i want to show the falsehood of what I take to be Skinner's central philosophical claim, on which all the others rest, and which he apparently derives from his vision of psychology. The claim is that *behavioral science proves that people are not free, dignified, morally responsible agents*

The first step in Skinner's argument is to characterize his enemy "mentalism". He has a strong gut intuition that the *traditional* way of talking about and explaining human behavior in "mentalistic" terms of a person's beliefs, desires etc. is somehow utterly disqualified

- Not only is it not science; it could not be turned into science; it is inimical to science, would have to be in conflict with any genuine science of human behavior

1. The most ancient and familiar worry about mentalism is the suspicion that mental things must be made of non-physical stuff, thus raising the familiar and apparently fatal problems of Cartesian interactionism (dualism)
2. Skinner rejects the common behaviorist claim that is the privacy of the mental in contrast to the public objectivity of the data of behavior that makes the mental so abhorrent to science
3. Skinner often inveighs against appealing to events whose occurrence "can only be inferred"
    - Chomsky takes this to be Skinner's prime objection against mentalistic psychology, but Skinner elsewhere is happy to note that "Science often talks about things it cannot see or measure" so it cannot be that simple
    - Skinner himself on occasion explicitly infers the existence of such events; it must be a particular sort of inferred events.
4. In particular *internal events* are decried, for they "have the effect of diverting attention from the external environment"

The first four reasons Skinner cites are all inconclusive or contradicted by himself

- Which I haven't really written why Dennett argues this is the case, but I probably should

If we are to go any further in characterizing Skinner's enemy we must read between the lines

Between the Lines

In several places Skinner hints that what is bothering him is the *ease* with which mentalistic explanations can be concocted

- One *invents* whatever mental events one needs to "explain" the behavior in question
- One falls back on the "miracle working mind" which just because it is miraculous "explains nothing at all"

Now this is an ancient and honorable objection vividly characterized by Moliere as the **virtus dormitiva**. The learned "doctor" in *Le Malade Imaginarie* on being asked to explain what it was in the opium that put people to sleep, cites its **virtus dormitiva** or sleep-producing power.

Leibniz similarly lampooned those who forged "expressly occult qualities or faculties which they imagined to be like little demons or goblins capable of producing unceremoniously that which is demanded, just as if watches marked the hours by a certain horodeictic (time-depending, red) faculty without having need of wheels, or as if mills crushed grains by a fractive faculty without needing any thing resembling millstones"

BUT...
Mentalistic explanations do not seem to cite a **virtus dormitiva**. For instance, explaining Tom's presence on the uptown bus by citing his desire to go to Macy's and his belief that Macy's uptown does not look like citing a **virtus dormitiva**: it is not as empty and question begging as citing a special uptown-bus-affinity in him would be.

There is a special case of virtus dormitiva in fact alluded to in the Leibniz passage, which is the key to Skinner's objection

- Sometimes the thing the desperate theoretician postulates takes the form of a little man in the machine, a homunculus

Skinner often alludes to this fellow: "The function of the inner man is to provide an explanation which will not be explained in turn"

This is a typical case of Skinner's exasperating habit of running together into a single undifferentiated lump a number of distinct factors that are related.

- Skinner sees superstition and demonology every time a claim is made on behalf of moral responsibility

In Skinner's eyes, psychologists who study physiology (and hence look at *internal* things) are scarcely distinguishable from folk who believe in witches

Skinner sees - or almost sees- that there is a special way that questions can be begged in psychology and this way is akin to introducing a homunculus. Since psychology's task is to account for the intelligence or rationality of men and animals, it cannot fulfill its task if anywhere along the line it *presupposes* intelligence or rationality. "...a homunculus is hiding in effect in your explanation *whenever you use a certain vocabulary* because the use of that vocabulary presupposes intelligence or rationality"

homunculus, presupposes intelligence or rationality. For instance, if I say that Tom is taking the uptown bus because he *wants* to go to Macy's and *believes* Macy's is uptown, my explanation of Tom's action *presupposes* Tom's intelligence, because if Tom weren't intelligent enough to put two and two together, as we say, he might fail to see that taking the uptown bus was a way of getting to Macy's. My explanation has a suppressed further premise: expanded it should read: Tom believes Macy's is uptown, and Tom wants to go to Macy's, so *since Tom is rational* Tom wants to go uptown, etc. Since I am relying on Tom's rationality to give me an explanation, it can hardly be an explanation of what makes Tom rational, even in part.

Dennett elaborates with an example of a chess playing computer *recognizing* what a good move is

Skinner comes very close to seeing this problem of presupposing intellect:

> "Nor can we escape.... by breaking the little man in to pieces and dealing with his wishes cognitions motives and so on, bit by bit. The objection is not that those things are mental but that they offer no real explanation and stand in the way of a more effective analysis"

These terms, the use of which presupposes the rationality of the entity under investigation, are what philosophers call the **intentional idioms**

Thus, spruced up, Skinner's position becomes the following: don't use intentional idioms in psychology

Spruced-up Skinner is not alone in being opposed to intentional idioms in psychology. His Harvard colleague, Quine, has been most explicit on the topic. (Quine as in Duhem-Quine and Quine - Two Dogmas of Empiricism)

- Quine's objections to intentional idioms are to be foresworn because we cannot translate sentences containing intentional idioms into sentences lacking them
- Skinner on the other hand is blithely confident that such translations are possible, and indeed *Beyond Freedom and Dignity* consists in large measure of samples of Skinner's translations

If Skinner never avails himself of the Chisholm-Quine untranslatability argument, and never makes explicit the presupposition of rationality argument, he does nevertheless muddy the water with a few other inconclusive objections

- Intentional explanations tend to be "unfinished", he says, in that an action is explained by reference to an opinion, without the existence of the opinion being explained in turn
- He also suggests that intentional explanations are not predictive, which is manifestly false
    - Tom/Macy bus example, which Dennett believes is pretty reliable

So far so good but Skinner again makes an important misstep by drawing the conclusion that, *intentional idioms therefore have no legitimate place in any psychological theory*

- There is not reason why intentional terms cannot be used provisionally in the effort to map out the functions of the behavior control system of men... just as long as a way is found eventually to "cash them out"
- Mendelian genetics is an example
    - "Mendelian genetics thrived as a science for years with nothing more to feed on that the concept of a gene, a whatever-it-turns-out-to-be that functions as a transmitter of a heritable trait"

It's this misstep that leads Skinner into his most pervasive confusion

- Skinner, unlike Quine, thinks that translation of intentional into non-intentional terms is possible
    - But why can't intentional explanations in virtue of these bonds of translation, find a spot in psychology?

Skinner vacillates between saying they can and they can't

> Beliefs, preferences, perceptions, needs, purposes, and opinions are possessions of autonomous man which are *said to* change when we change minds. What *is* changed in each case is a probability of action. (*my italics*)[32]

> How are we to interpret this? As meaning that we change probabilities, *not* beliefs, or as meaning that changing beliefs *is just* changing probabilities of action? Skinner's very next sentence strongly suggests the latter:
>
> A person's belief that the floor will hold him as he walks across it depends upon his past experience.

> but a few sentences later he hedges this by putting "belief" in scare-quotes:
>
> We build "belief" when we increase the probability of action by reinforcing behavior.
>
> Does this passage mean that it is *all right* to talk of building belief, so long as we understand it as increasing action probabilities, or that it is

> *wrong* to talk that way since *all* we are doing is increasing action probabilities?[33] On the next page he takes the hard line:
>
> We change the relative strengths of responses by differential reinforcement of alternative courses of action; we do *not* change something called a preference. We change the probability of an act by changing a condition of deprivation or aversive stimulation; we do *not* change a need. We reinforce behavior in particular ways; we do *not* give a person a purpose or an intention. (*my italics*)

Skinner is apparently quite well versed in vacillation

- Dennett presents a bunch of examples of Skinner switching opinions

Dennett argues that intentional and mechanistic or scientific explanations can co-exist

- He gives the computer playing chess example, which eventually boils down from "recognizes moves" to "it's simply at the mercy of electronic currents"
- What this chain adds up to "prove" is that computers are just dull lumps of stuff, they can't do anything
  - But of course they can. What the PC programmer can do if we give him the chance is not *explain away* the illusion that the computer is doing these things, but *explain how* the computer is truly doing them
- Skinner fails to see the difference between explaining and explaining away

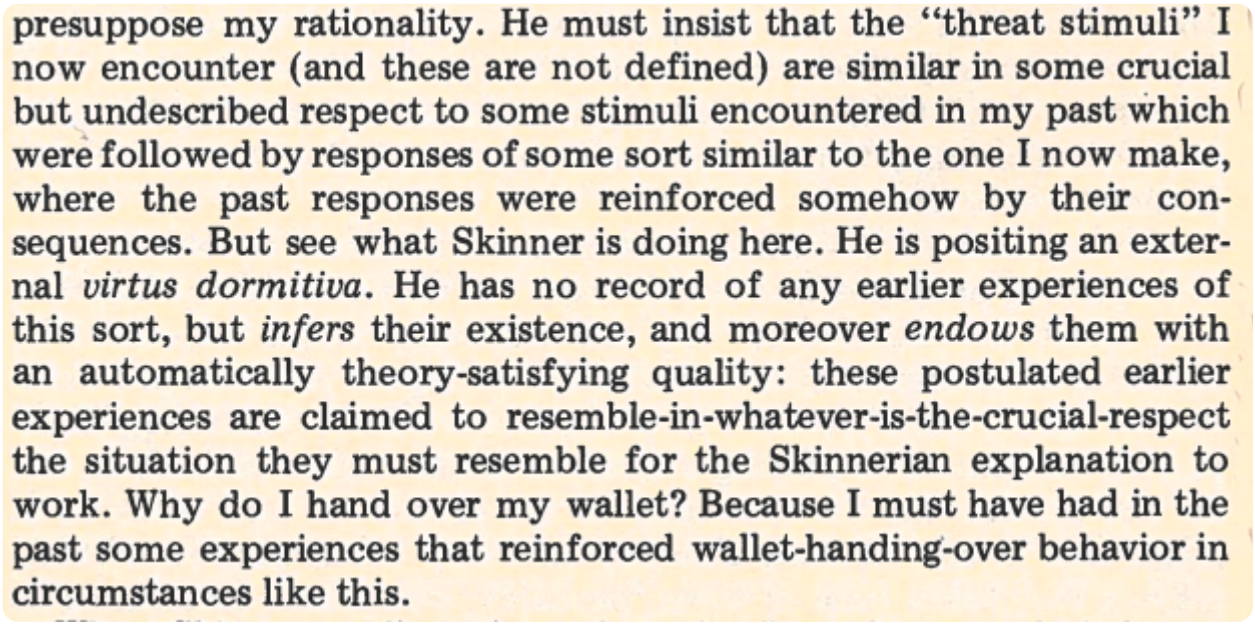
    - In this regard he is succumbing to the same confusion as those who suppose that since color can be explained in terms of the properties of atoms which are not colored, nothing is colored
- Since Skinner fails to make this distinction, he is led to the exclusive view, the view that true scientific explanations will exclude true intentional explanations, and typically, he offers no arguments for it

The fact that it is the simplicity of explanations that can render elaborate intentional explanations false is completely lost to Skinner for a very good reason: the only *well-formulated, testable* explanations Skinner has so far come up with have been, relatively simple, and deal with relatively simple behavior controls of relatively simple animals.

The persistently recalcitrant features of human behavior for the Skinnerians can be grouped under the headings of novelty and generality. The Skinnerian must explain all behavior by citing the subject's history if similar stimuli and responses, so when someone behaves in a novel manner, there is a problem.

- I am held up (threatened) and asked for my wallet; this has never happened to me before so the correct response cannot have been "reinforced" for me, yet I do the smart thing: I hand over my wallet
- The Skinnerians must claim that this is not truly novel behavior at all, but an instance of a*general sort* of behavior which has been previously conditioned.
- It is perfectly clear that what experience has taught me is that if I *want* to save my skin, and *believe* I am threatened, I should do what I *believe* my threatener *wants* me to do.
    - Of course Skinner cannot permit this intentional formulation at all, for in ascribing wants and beliefs it would presuppose my rationality

Further elaboration:



presuppose my rationality. He must insist that the "threat stimuli" I now encounter (and these are not defined) are similar in some crucial but undescribed respect to some stimuli encountered in my past which were followed by responses of some sort similar to the one I now make, where the past responses were reinforced somehow by their consequences. But see what Skinner is doing here. He is positing an external *virtus dormitiva*. He has no record of any earlier experiences of this sort, but *infers* their existence, and moreover *endows* them with an automatically theory-satisfying quality: these postulated earlier experiences are claimed to resemble-in-whatever-is-the-crucial-respect the situation they must resemble for the Skinnerian explanation to work. Why do I hand over my wallet? Because I must have had in the past some experiences that reinforced wallet-handing-over behavior in circumstances like this.

Skinner can't explain complex human behavior, and relies increasingly on a **virtus dormitiva**

Reading Questions

Dennett:

1. On p. 56, Dennett refers to *virtus dormitiva* in a play by Molière: "The learned "doctor" in *Le Malade Imaginaire*, on being asked to explain what it was is the opium that put people to sleep, cites its *virtus dormitiva* or sleep-producing power". What aspects of behaviourism may also be vulnerable to this type of criticism according to Dennett?
2. Are *internalist* explanations (using mentalistic concepts such as "want", "desire", "believe" etc.) useful, according to Dennett, in explaining novel behavior?
3. If they are, what are the implications for behaviorism?

Answers:

1. Skinner is incapable of explaining complex behavior (e.g. human's exhibiting novel behavior in novel situations), which forces him (in Dennett's words) to attribute this to *virtus dormitiva*
2. Yes, they are crucial in fact.
3. They serve as one of the main arguments against radical behaviorism

Hampton - Metacognition in Nonhumans

Robert R. Hampton

Abstract:
I review representative examples from this literature, considering the sufficiency of four classes of mechanism to account for the metacognitive performance observed. This analysis suggests that many of the demonstrations of metacognition in nonhumans can be explained in terms of associative learning or other mechanisms that do not require invoking introspection or access to private mental states

Metacognition: A broad definition

Metacognition allows one to monitor and adaptively control cognitive processing

For example, a student might improve her grade by dedicating more of her study effort to the longest textbook chapters and the most difficult topics on an upcoming exam.

Our student's behavior in each of the above examples deserves the label metacognition, at least as it is broadly defined (Flavell, 1979). Demonstrations of metacognition in the laboratory must meet four criteria:

1. We must specify a primary, objectively observable behavior that can be scored for accuracy or efficiency. Accuracy might be assessed as questions answered correctly, while efficiency could be assessed as time taken to learn all the assigned material.
2. There must be variation in the accuracy or efficiency of the primary behavior. Variation in performance is necessary in order to allow assessment of the correlation between the primary behavior and the secondary behavior (described in 3, below).
3. We must specify a secondary, objectively observable behavior that can be used to infer monitoring or regulation of cognition underlying the primary behavior. Monitoring of knowledge might be indicated by skipping questions for which the subject is unsure of the answer, while regulation might be indicated by subjects adjusting time spent studying to match the difficulty of the material.
4. There must be an explicit assessment of whether the primary and secondary behaviors are correlated. For example, were the questions that the subject

skipped indeed ones for which he did not know the answer? Was study time adjusted appropriately to increase efficiency of learning?

Given the objective nature of the four criteria outlined above, it is possible to devise tests of metacognition for nonhuman animals. In fact, a substantial literature has developed over the last 15 years demonstrating that several nonhuman species clearly meet all four of these criteria.

- In perceptual tests, monkeys, dolphins, and rats have been shown to either decline difficult trials or make accurate post-trial confidence judgments

Private and public mechanisms for metacognition

Metacognition in humans is often associated with conscious awareness of one's own cognitive states and is therefore presumed to reflect private monitoring of those states

Private mechanisms are those by which cognitive control is contingent on the privileged access the subject has to their own cognitive states.
In the case of *public* mechanisms, adaptive cognitive control is based upon the use of publicly available information such as the perceivable difficulty of a problem or the subject's reinforcement history with particular stimuli.

Contrast the following two situations requiring a metacognitive judgment:

1. a colleague asks whether you remember the title of B. F. Skinner's first book,
2. a friend asks whether you can answer a question his six year old has about psychology

In the first case, you would surely check the contents of your memory and determine whether you can retrieve a memory of the book title. Your metacognitive judgment would therefore depend on your success or failure at privately retrieving the relevant explicit memory, a cognitive state to which you, as the one doing the remembering, have privileged access.

In the second case, your friend has not even asked you to retrieve a specific memory. If you are an expert in Psychology, you might feel confident (probably correctly) that you can answer the question of a six year old. However, your confidence would not depend on a private evaluation of your memory. Instead, your confidence would depend on your history of expertise, your past ability to answer such questions, and your assessment of the intellectual capacity of six year olds – all publicly available information

To understand the mechanisms of metacognition in nonhumans we will have to do more than demonstrate adaptive cognitive control.

- We will have to develop experimental procedures that allow us to specify what information subjects use to assess their ability to learn or perform, and how they use that information

A fear probably shared by all investigators of nonhuman metacognition is that we are misinterpreting "*Clever Hans*" type phenomena in which apparently impressive cognitive feats can be accomplished by established "simple" mechanisms
Humans can accomplish a great deal of learning and performing without conscious awareness or introspection, for example,

- classical conditioning, skill learning, and priming

Because public mechanisms of metacognition depend on publicly observable information, their operation can likely be explained in terms traditional to animal learning and comparative cognition. By contrast, evidence for private mechanisms involving some type of introspection might require that we extend our understanding of what nonhuman animals perceive to include some of their own cognitive states.

Four classes of stimuli sufficient for metacognitive control

Most or all cases of nonhuman metacognition may be adequately accounted for by public mechanisms.

Because we cannot obtain from nonhumans the verbal reports that constitute part of the evidence for private introspective metacognition in humans, we can only infer private metacognition in nonhumans by excluding likely public mechanisms.

Environmental cue associations

Some stimuli are more difficult to discriminate or remember than are others and some test conditions are more challenging than are others

- Stimuli that are close together on a continuum are more difficult to discriminate than are those that are far apart.
- Highly similar images are difficult to identify in matching-to-sample tests.
- etc.

Subjects performing tests with such stimuli might use the identity, magnitude, similarity, delay, or other publicly available information as a discriminative cue for declining tests or rating confidence. For example, if subjects have experienced low rates of reward

with stimuli in a specific magnitude range, they could learn to avoid tests with all stimuli in that range

The probability that *Environmental Cue Associations* can account for performance in a given paradigm is best assessed by generalization tests which determine whether or not performance is maintained across changes in the particular stimuli used and specific conditions of testing.

- If performance immediately generalizes to new test conditions or new stimuli, it is safe to conclude that metacognitive responding was not controlled by stimuli that were changed for the generalization test.

Behavioral cue associations

Similar to the aforementioned class with the exception that the discriminative stimuli controlling use of the metacognitive response are systematically generated by the subject in a way that correlates with accuracy in the primary task

- For example, the subject may vacillate when it does not know the correct response on a given test

This vacillation does not necessarily represent metacognition by the subject that it does not know the answer, but can rather be an unmediated result of not knowing how to respond. It is common to see this sort of vacillation in monkeys taking matching-to-sample tests, for example, in which they look back and forth between the choice stimuli before choosing (personal observations).

- It is also well known that response latency is often longer for incorrect than correct responses.
- Because vacillation and response latency correlate with accuracy, subjects could use these self-generated cues as discriminative stimuli for the metacognitive response, for example by declining tests on which they experience a relatively long response latency

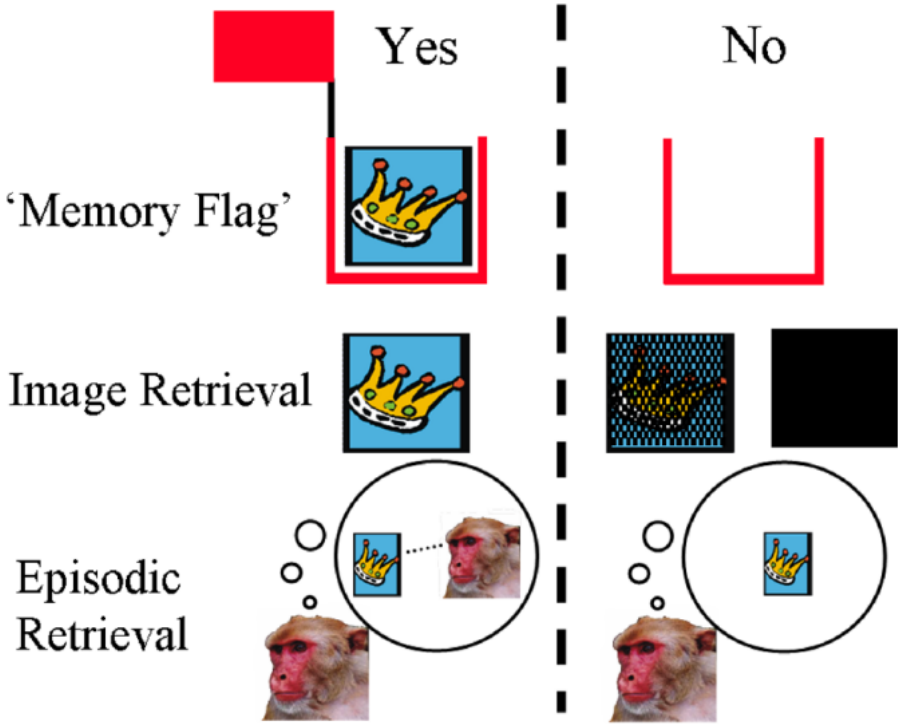
Response competition

In most reports of metacognition in nonhumans, subjects are confronted with the primary discrimination problem or memory test and the secondary metacognitive response option simultaneously. Because subjects can only make one response (a primary test response or a secondary decline test response, for example), simultaneous presentation puts these two behaviors in direct competition. *Response Competition* can be ruled out as an account for metacognitive responding by presenting the secondary metacognitive response option either before or after the primary test, so that the two types of response do not compete directly.

Introspection

Metacognition could also be mediated by private introspective assessment of the subject's mental states. While introspection might not necessarily require consciousness, it is closely allied with consciousness in humans

By the introspection account, the discriminative stimulus controlling a metacognitive response (e.g. declining to take a test) is the private experience of uncertainty or the weakness of memory.). In the case of uncertainty, subjects are suggested to experience conscious (at least in humans) "feelings of uncertainty" that differ from the experience of objective stimuli.
In the case of memory, subjects are proposed to assess the strength of their memory. The assessment of memory might be accomplished through several mechanisms that vary in sophistication from detecting whether a memory is present (while knowing nothing of the content of the memory) to attempting to retrieve the relevant memory and determining the success of that effort (Figure 1).



Cartoons of three candidate processes of introspective memory assessment.

- The column to the left of the dashed line represents the contents of cognitive processing on trials on which monkeys choose to take the memory test.
- The right column depicts the same on trials on which the test is declined. The memory flag hypothesis posits an indicator for the presence of memory. Monkeys use the metacognitive response contingent on the indicator, but are not aware of the content of the memory.
- In the case of the image retrieval model, the decision to take the memory test is based on the vividness of the memory retrieved. The episodic retrieval hypothesis

proposes that monkeys take the test when they can remember the context of the study episode and decline the test when this information cannot be recovered.

The important difference between this account and the preceding three is that use of the metacognitive response is based on privileged introspective access to the subject's cognitive states, rather than on publicly available information or Response Competition. Due to the private nature of *Introspection*, the conclusion that it accounts for metacognitive performance in nonhumans can probably be reached only by ruling out other accounts.

Evaluation of the literature through selected examples

	Distinctive features	Environ-mental cues	Behavioral cues	Response competition	Intro-spection
Call & Capenter, 2001; Hampton, Zivin & Murray, 2004	opaque tubes, "spontaneous" metacognition	unlikely, counter-balanced stimuli	unlikely, limited experience	possible, concurrent responses	possible
Foote and Crystal, 2007	prospective tests, temporal psychophysics	possible	unlikely, prospective judgment	no, prospective judgment	possible
Hampton, 2001	prospective judgment, generalization tests	no, generalization to no-sample trials, delays	unlikely, prospective judgment, generalization to no-sample trials	no, prospective judgment	possible
Kornell, Son & Terrace, 2006	retrospective confidence judgment by "gambling"	no, generalized to new stimuli	possible, response latency	no, retrospective judgment	possible
Smith, Schull, Strote, McGee, Egnor & Erb, 1995	dolphin auditory psychophysics	possible	possible	possible	possible
Smith, Shields, Schull & Washburn, 1997	psychophysical pixel density test	possible	possible	possible	possible
Smith, Shields, Washburn & Allendoerfer, 1998	serial position effect	no, list position not perceptible at test	possible	possible	possible

Table 1. Characterization of selected experiments with respect to four classes of stimulus control for metacognitive responding.

- A green background indicates that the type of stimulus control indicated in the column heading can account for the reported metacognitive performance.

- A medium red background indicates that the indicated stimulus control is ruled out.
- A light yellow background indicates a low probability of stimulus control.
- Text indicates how a particular type of stimulus control was ruled out. To the extent that particular sources of stimulus control can be ruled out, the remaining sources of control are more likely to be in effect

Dolphin Auditory Psychophysics (J. D. Smith, et al., 1995).

The first report of metacognition in a nonhuman species described the performance of a bottlenosed dolphin (Tursiops truncatus) in an auditory psychophysical task.
This publication nicely introduces many of the major features common to tests of nonhuman metacognition

A dolphin was trained to distinguish between tones of 2100 Hz and tones below it. Initially, it learned to respond to a left paddle for 2100 Hz tones and a right paddle for lower frequencies. As the frequency neared 2100 Hz, its accuracy dropped, often mistaking close tones for 2100 Hz.
After mastering this, a third paddle was added, allowing it to skip difficult trials in favor of an easy one (1200 Hz). The dolphin learned to maximize rewards by choosing the primary discrimination for easy trials and the third paddle for difficult ones. Its behavior aligned with these strategies, rarely using the third paddle for easy trials but increasingly using it for near-2100 Hz tones, the toughest trials.

The dolphin clearly met the criteria for metacognition, adaptively taking easy tests and declining difficult tests. How might the dolphin have accomplished this?

Several features of this experiment suggest that the dolphin may have used publicly observable cues to guide use of the decline response, suggesting *public metacognition*

First, the dolphin may have used tone frequency as a discriminative stimulus for making a decline response to the third paddle. In this design, discrimination difficulty is confounded with frequency, that is, difficulty and frequency are correlated. The dolphin may have learned to select the decline test paddle in the presence of stimuli belonging to a particular frequency range because of its reinforcement history with those particular tones, rather than because of a subjective feeling of uncertainty. Thus, Environmental Cue Associations may be sufficient to account for the metacognitive performance.

- This account could be tested by determining whether use of the third paddle response generalized immediately to tests conducted in new frequency ranges (e.g., 3100 Hz vs. 2200- 3099 Hz).

    - If the dolphin had learned a general metacognitive response, it should continue to avoid difficult trials in the new frequency range
- By contrast, if the dolphin had learned to use the decline test response whenever tones of specific stimuli were used, the dolphin would have to relearn which frequencies should occasion this response through trial and error learning of which frequencies were associated with low rates of reward

Second, the dolphin may have used its own publicly observable behavior as a discriminative stimulus for declining tests. As described earlier, subjects often *vacillate* on difficult trials, a pattern also reported for the dolphin as "ancillary behaviors" near threshold.
It is tempting to interpret these "ancillary behaviors" metacognitively, as indicating that the subject hesitates because it knows it is uncertain. However, it is safer to interpret them non-metacognitively; when the subject does not know how to respond, it is slow to do so and may engage in other behavior in the meantime.

- Thus, the dolphin may have learned to use its own vacillation as a discriminative stimulus for the decline response, a type of **Behavioral Cue Association**.

Third, in this experiment (and many others) the secondary metacognitive response and the primary choice response were presented simultaneously, admitting the possibility that **Response Competition** can account for metacognitive performance. Simultaneous presentation places selection of one of the primary test responses (left or right paddle) in direct conflict with selection of the secondary decline response (the third paddle).

Finally, the dolphin may have used Introspection, or private metacognition. By this account, the dolphin reacted to a private cognitive discriminative stimulus (e.g. subjective uncertainty) that indicated that it did not know the correct answer on specific trials. Because multiple public accounts are viable, invoking an introspective account may be unwarranted.

Collecting Information When Ignorant (Basile, et al., 2009; Call & Carpenter, 2001; Hampton, et al., 2004)

Metacognition is evident when subjects collect additional information when ignorant and act immediately when informed.
Call and Carpenter (2001) developed a clever test of this capacity and used it with human children, chimpanzees (Pan troglodytes), and orangutans (Pongo pygmaeus). A modified version of this same test was subsequently used with rhesus monkeys (Macaca mulatta, Hampton, et al., 2004) and capuchin monkeys (Cebus apella, Basile, et al., 2009).

Subjects were presented with a set of opaque tubes in which food was hidden. Subjects either witnessed the baiting (seen trials) or did not (unseen trials), and therefore were either informed or ignorant about the food's location on each trial. At test, subjects could select a single tube and collect the reward, if correct. This test is an interesting assessment of metacognition because the subjects could bend over and look down the length of the tubes to locate the food before choosing (see 2nd Semester/Philosophy of Cognitive Science/7 Behaviorism/Attachments/Figure 2.png).

- Subjects demonstrate metacognition by collecting information when ignorant (unseen trials) and choosing immediately when informed (seen trials).
    - Human children, chimpanzees, orangutans, and rhesus monkeys clearly showed this pattern of behavior, while the case for capuchin monkeys was less clear

How does this performance relate to the four accounts of metacognitive behavior under consideration? This discussion will focus on the representative study with rhesus monkey subjects (Hampton, et al., 2004)

Unlike with the case of the dolphin psychophysical test, it is not possible that the rhesus used the identity of the test stimuli to guide their decision to look because each tube was equally likely to contain the food on both seen and unseen trials.

- Monkeys were familiarized with the apparatus and procedures (including gaining experience with looking down tubes) in such a way as to prevent them from learning via differential reinforcement to look selectively on unseen trials.
- Furthermore, comparatively few critical test trials were presented, to prevent monkeys from developing associations between experimental cues and the probability of reinforcement

It is therefore unlikely that **Environmental Cue Associations** or **Behavioral Cue Associations** underlie performance in these tests
These tests, like many other tests of metacognition, presented the primary choice response and the secondary metacognitive response *simultaneously*

- Knowing the food's location may strongly predispose a monkey to select that tube, decreasing the occurrence of all other possible behaviors, including searching the tubes
- Consequently, the metacognitive performance of monkeys in this paradigm may be the result of **Response Competition.**

Finally, monkeys may have used *Introspection* however, as the behavior can be explained by at least one public mechanism (Response Competition), more research is

needed d before we can safely infer a private mechanism

Serial Position and Confidence about Memory (J.D. Smith, et al., 1998)

When subjects are presented with lists of items to remember it is typical for items early and late in the list to be remembered better than middle items. Smith et al. (1998) took advantage of this predictable pattern of memory strength to assess whether monkeys showed metacognition for memory in a serial probe recognition task.

Monkeys saw a list of four consecutive random dot polygon figures and their memory for individual polygons from the list was probed using a yes-no recognition test. Monkeys showed the expected serial position effect; their memory was better for the first and last items than for the middle items.
Monkeys were then presented with a decline test response, similar to that used with the dolphin psychophysical test, concurrently with a probe polygon that may or may not have been from the studied list. Consistent with adaptive metacognitive control, the monkeys declined tests of the middle list items more often than tests of the first and last list items.

Because no publicly observable aspect of the test setup correlated with memory strength, it is unlikely that the metacognitive response was under control by **Environmental Cue Associations.**

Because the monkeys had ample opportunities during training to associate long response latencies with sparser rewards following the primary choice response, metacognitive performance could be under the control of **Behavioral Cue Associations.** (There's more but can't be bothered.....) on pg. 8 if wanna read

Like all examples discussed thus far, the primary yes-no recognition test and the secondary decline response were presented simultaneously. Thus, **Response Competition** could account for the observed metacognitive performance.
Finally, **Introspection** also remains a viable account of metacognitive performance in this study. However, we must again be cautious in inferring Introspection until we can rule out possible public mechanisms.

Retrospective Metacognitive Judgments

Probably the most creative of the published nonhuman metacognition paradigms is the retrospective *gambling* paradigm (Kornell, et al., 2007; Son & Kornell, 2005)

In this paradigm, monkeys rated their “confidence” by wagering either a large or small number of video tokens on the accuracy of each test trial immediately after they completed it. The video tokens were secondary reinforcers that were periodically

“cashed out” for actual food when a sufficient number had accumulated. Critically, monkeys placed their wager after answering, but before receiving feedback about their accuracy.
In this paradigm, metacognition predicts large wagers following easy tests and small wagers following difficult test.

- This is indeed how the monkeys performed, suggesting that they knew whether they had responded correctly despite the lack of feedback prior to placing their bet.

Presentation of the metacognitive response after completion of test trials effectively rules out **Response Competition** as a viable account for metacognitive performance; that is, performing the primary test response does not directly lower the probability of performing the secondary metacognitive response.

Kornell et al. (2007) also ruled out **Environmental Cue Associations** by showing that use of the metacognitive gambling response generalized across stimuli and, more importantly, across test types (from perceptual tests to a mnemonic test).

Behavioral Cue Associations remain a potential source of metacognitive control in these studies. Although separating the secondary metacognitive response from the primary task is a powerful control procedure, offering the metacognitive response *after* the primary test means that the subjects have already directly experienced the difficulty of each trial before they have to make their wager.

Finally, **Introspection** also remains a viable basis for control of the metacognitive response in these studies. Future studies might focus on ruling out stimulus control by response latency.

Prospective Metacognitive Judgments

Bunch of text

Converging evidence?

The reviewed studies show that animals adaptively regulate decisions about when to take tests, when to collect more information, and how to rate their own performance. It is much less clear whether these studies provide converging evidence regarding the mechanisms by which this adaptive cognitive control is achieved.

Inspection of Table 1 shows that **Introspection** is always a potential source of stimulus control in these studies; *however*, even consideration of this limited set of possible alternative accounts shows that metacognitive control in most studies can be adequately explained without invoking introspection.

Thus, there is a high bar to clear in terms of ruling out alternative mechanisms for metacognition before we can conclude that any nonhuman animals engage in private metacognition.

What are we trying to learn? Relationship to implicit and explicit representation

It may still be premature to conclude that any case of observed metacognition in nonhumans depends on introspection involving explicit representations, but when sources of public stimulus control are eliminated, it is more likely that Introspection underlies metacognitive performance

Implications for Comparative Psychology

It is intriguing that it appears to be easier to demonstrate metacognition in some species than in others. For example, while there are many reports of metacognition in rhesus monkeys, work with pigeons has been much less likely to detect metacognition. In parallel tests conducted with human children, apes, rhesus monkeys, and capuchin monkeys, capuchin monkeys show by far the weakest evidence for metacognition.

It is tempting to interpret these differences as indicating that metacognitive control is not “easy,” is unlikely to come about through “simple” associative learning (of which pigeons and capuchins are certainly capable), and may be restricted to relatively few species.
However, it is still too early to reach this conclusion; there are a host of species characteristics that may interfere with performance in metacognitive tests (e.g. differences in *attention*, *impulse control*, and *motivation*). Sorting out which species can and cannot behave metacognitively will be greatly helped if we can agree as a community what behavioral criteria are required.

Reading Questions

Hampton:

1. How does Morgan's Canon (C. Lloyd Morgan) relate to Hampton's efforts of establishing that monkeys might rely on introspection for certain tasks? The canon is thus stated: *In no case is an animal activity to be interpreted in terms of higher psychological processes if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development.*
2. Can we exchange "animal" for "human" in the formulation above

Answers:

1. It's a pretty strong counter-argument if true, because all of the cases/experiments that Hampton outlines have lower psychological processes that could provide basis for interpretation, thus ruling out the possibility of Introspection
2. You could yeah, although we have free will so, yeah. Kind off an open-ended question wouldn't you say?

Computationalism Lecture

The overview:

- Cognition is computation on representations
- Computation is implementation independent (it does not what hardware it is run on, e.g. brains or computers)
- Computation is systematically interpretable symbol manipulation (it has strict syntax)

Learning goals

1. Understanding the intellectual history leading to the **cognitive revolution**
2. Understanding the **assumptions of computationalism**
3. Understanding what (mental) **representations** are
4. Understanding David Marr's three levels of analysis
    1. see also Marr - Vision, A Computational Investigation

The Second Overview

- Philosophy of science
    - Computationalism: a paradigm shift
- Historical background
    - Emergence of cognitive science
- Computationalist cognitive science

- Cognition = Computation
- From psychology to cognitive science

What are Paradigms again (good old Kuhn - The Structure of Scientific Revolutions)

- The practices that define a scientific discipline
  - What phenomena do we look at?
  - Which questions do we ask?
  - How do we design experiments?
  - How do we interpret results?

The behaviourist paradigm

- The practices that define a scientific discipline
  - What phenomena?
  - Which questions do we ask?
  - How do we design experiments?
  - How do we interpret results?
- Behaviourist practices
  - Operants, responses reinforcement, punishment
  - How do schedules of reinforcement shape behaviour?
  - Control the environment and observe frequency of behaviour
  - Without invoking inner causes (externalism)

Computationalism: A paradigm shift

- Computationalist practices
  - Range of behaviour made possible by internal computation
  - Which computational processes underlie this range of behaviour?
  - Test the flexibility and constraints of cognition
  - Infer the inner computational operations

Reflection point: Do old paradigms truly disappear?

- Not necessarily
  - Most animal studies use *behaviourist* methodologies to “instruct” their animals, i.e. they learn a task by operant conditioning

Morgan’s canon:

> “In no case is an animal activity to be interpreted in terms of higher psychological processes if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development.”

In plain words: only when certain capabilities cannot be explained behaviourally, can we invoke mental explanations (in animals)

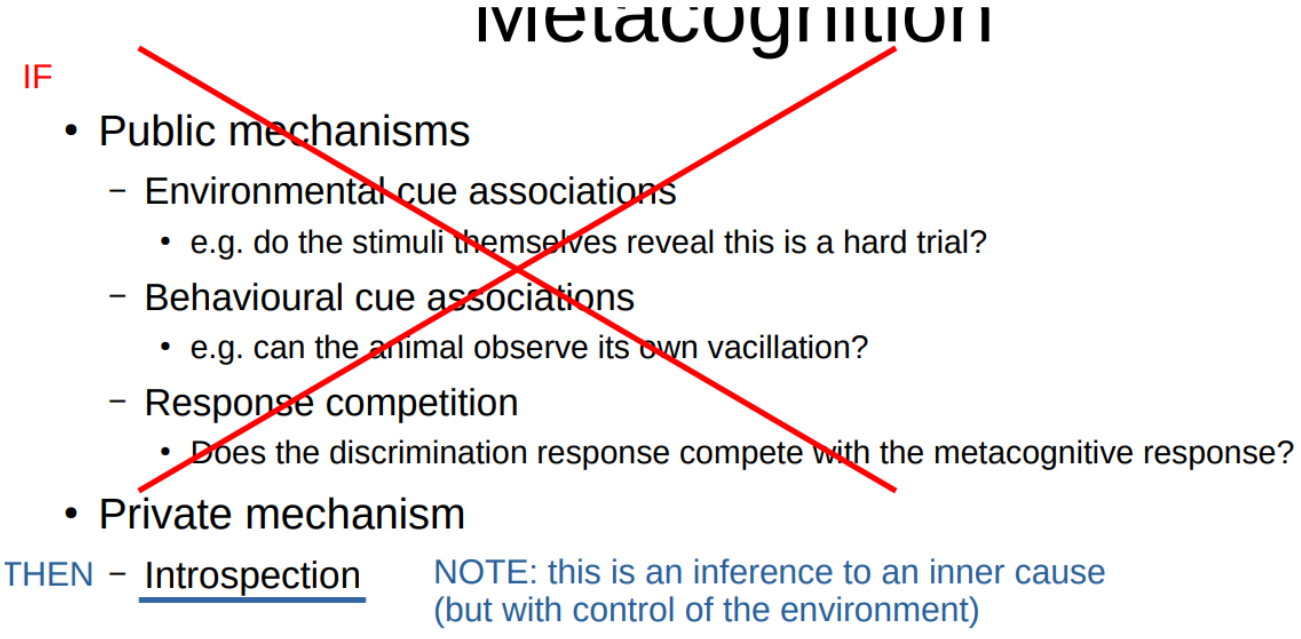
Case in point: **metacognition**: (knowing what you know and that you know)

Reference back to the Hampton - Metacognition in Nonhumans

- Do monkeys know that they do not know?
- More specifically do they show signs of metacognition?

Metacognition

- Public mechanisms
  - Environmental cue associations
    - e.g. do the stimuli themselves reveal this is a hard trial?
  - Behavioural cue associations
    - e.g. can the animal observe its own vacillation?
  - Response competition
    - Does the discrimination response compete with the metacognitive response?
  - Private mechanism
    - Introspection



- Basically, behavior should only be attributed to private mechanisms (introspection) if no other public mechanisms can be said to be responsible

Lau's personal take

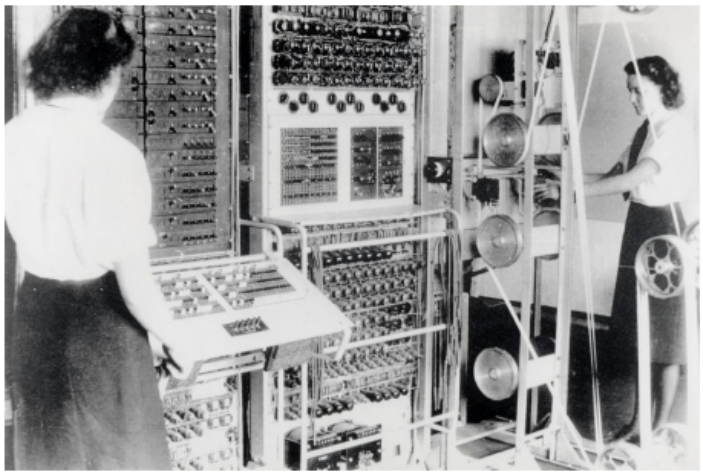
- don't make fanciful mental explanations, if behaviourist explanations will do!
- I believe he said the same thing in the Behaviorism Lecture
  - he did lol
    - Behaviorism Lecture > ^69814c

Questions About Paradigm Change

General: why do they come about at all?
Specific: why did computationalism come about?

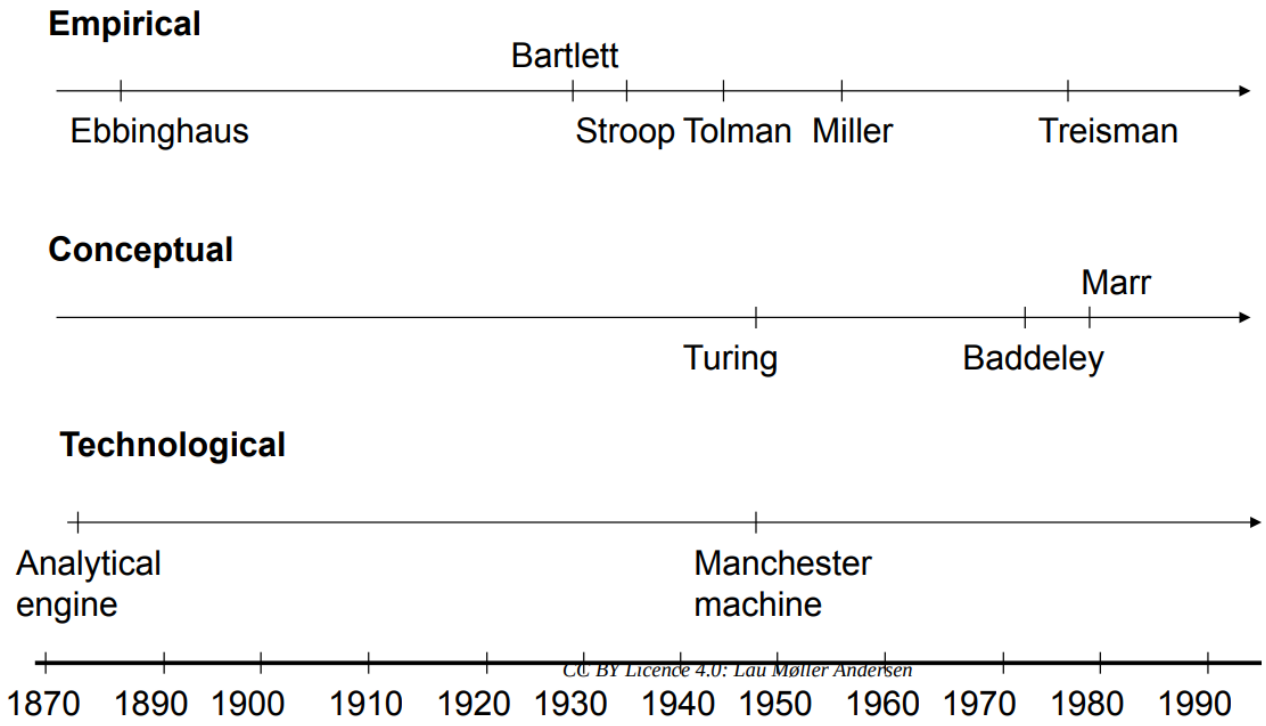
Why do they come about

- Empirical tension
  - body of data irreconcilable with the current paradigm
- Conceptual tension
  - concepts that are "forbidden" seem necessary
- Technological tension
  - new technologies suggest new metaphors



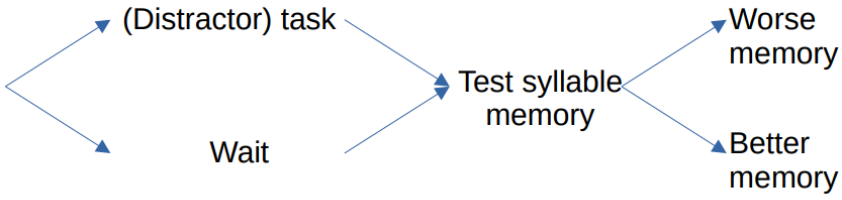
Colossus, the first electronic digital programmable computing device, was used to break German ciphers during World War II. It is seen here in use at Bletchley Park in 1943. https://en.wikipedia.org/wiki/Computer#/media/File:Colossus.jpg

TENSIONS (not an exhaustive list)



Ebbinghaus 1885, internal processing

bak	cug	dep	fik
gom	haf	jun	kip
laf	muv	nop	pud
ret	sal	tiv	wug



An interpretation: distractor task interferes with internal memory processes

Studies of **interference** in **serial verbal reactions**.
JR Stroop - Journal of experimental psychology, 1935 - psycnet.apa.org
In this study pairs of conflicting stimuli, both being inherent aspects of the same symbols, were presented simultaneously (a name of one color printed in the ink of another color—a word …
☆ Save 99 Cite Cited by 25938 Related articles All 20 versions Web of Science: 13099

McGurk effect
Stroop interference
Reasoning errors
Visual Illusions
Recency / primacy effects
Pop-out effect
Many priming effects

Green **XXX**
Red **XXX**

An interpretation: Some processes are automatic, others require (metacognitive) control

Slide adapted from Iris van Rooij

Turing and his Machine

On computable numbers, with an application to the Entscheidungsproblem
(14k citations btw)

Turing machine: An abstract “machine” that can perform any algorithm that a human could execute

Input: symbols
Operations: formal
Output: processed symbols

Icons, indices (index), and symbols:

Icon: ***represents*** **through similarity (intrinsic relation)**

	Index: *represents* by pointing to (the fire I'm guessing)
A	Symbol: *represents* by convention or definition (no intrinsic relation)

Many symbols are ambiguous: "A" maps to several sounds for example

Systematically interpretable

INTERPRETATIONS DON'T CHANGE

```
interpretations = (
                   (True,  True ),
                   (True,  False),
                   (False, True ),
                   (False, False)
                  )

for interpretation in interpretations:
    print('When argument 0 is ' + str(interpretation[0]) + \
          ' and argument 1 is ' + str(interpretation[1]) + ' then: ')

    if interpretation[0] and interpretation[1]:
        print('The "and" operator evaluates to "True"')
    else:
        print('The "and" operator evaluates to "False"')
    if interpretation[0] or interpretation[1]:
        print('The "or" operator evaluates to "True"')
    else:
        print('The "or" operator evaluates to "False"')
    if not interpretation[0] or interpretation[1]:
        print('The "if-then" operator evaluates to "True"')
    else:
        print('The "if then" operator evaluates to "False"')

    print('\n')
```

```
When argument 0 is True and argument 1 is True then:
The "and" operator evaluates to "True"
The "or" operator evaluates to "True"
The "if-then" operator evaluates to "True"


When argument 0 is True and argument 1 is False then:
The "and" operator evaluates to "False"
The "or" operator evaluates to "True"
The "if then" operator evaluates to "False"


When argument 0 is False and argument 1 is True then:
The "and" operator evaluates to "False"
The "or" operator evaluates to "True"
The "if-then" operator evaluates to "True"


When argument 0 is False and argument 1 is False then:
The "and" operator evaluates to "False"
The "or" operator evaluates to "False"
The "if-then" operator evaluates to "True"
```

... and must follow the machine syntax

```
for interpretation in interpretations:
    print('When argument 0 is ' + str(interpretation[0]) + \
          ' and argument 1 is ' + str(interpretation[1]) + ' then: ')

    if interpretation[0] og interpretation[1]:
        print('The "and" operator evaluates to "True"')
    else:
        print('The "and" operator evaluates to "False"')
    if interpretation[0] or interpretation[1]:
        print('The "or" operator evaluates to "True"')
    else:
        print('The "or" operator evaluates to "False"')
    if not interpretation[0] or interpretation[1]:
        print('The "if-then" operator evaluates to "True"')
    else:
        print('The "if then" operator evaluates to "False"')

    print('\n')
```

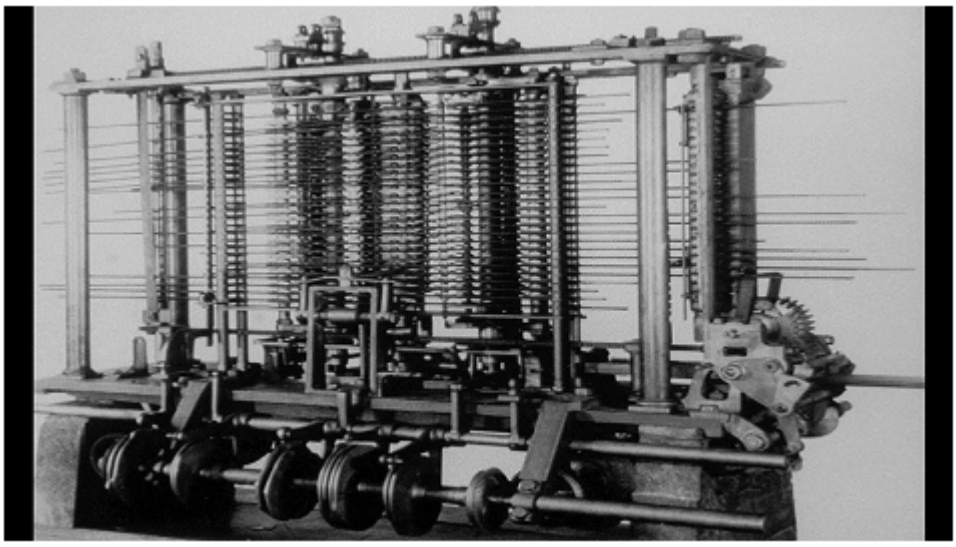
```
    if interpretation[0] og interpretation[1]:
                        ^
SyntaxError: invalid syntax
```

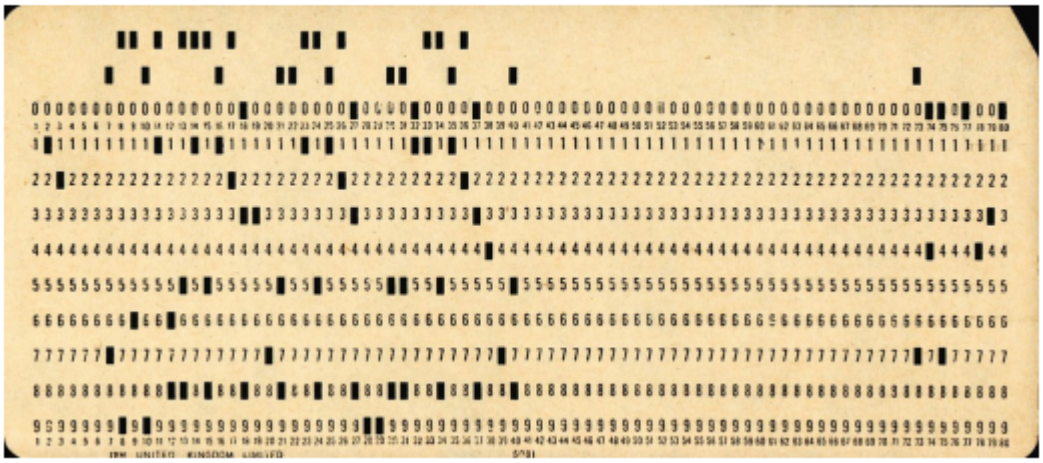
Mental representations are symbolic representations

Analytic Engine

- Steam powered loom
- Inputs and outputs of calculations represented as mechanical "loom" states
- The logic carrying inputs to outputs "programmed" on metal punch cards
- Machine calculates according to the punch cards' shape – logic is embedded in the shape
- Machine calculation is the manipulation of symbols according to logical rules

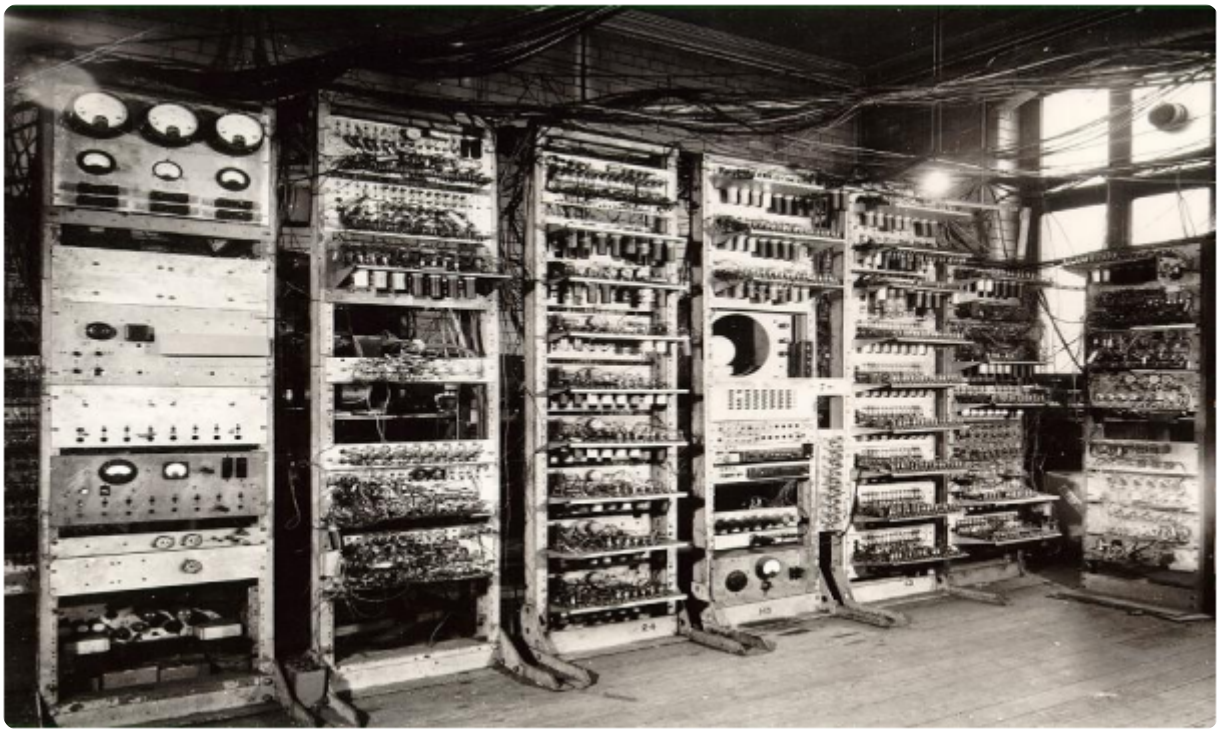
Correct to 6 decimal places!





The Manchester Machine - the first computer!!!

- Williams tube
  - CRT divided into an array
  - First instance of RAM
- Magnetic drum
  - Programmable magnetic tracks
  - Data input
  - Magnetic variation “represents” numbers
  - Punchcard, keyboard, or “tape”
- 4200 vacuum tubes
  - Logical processes: what to do with information in different parts of the memory
  - Logic gates or CPU



Lessons from early computer technology

1. Computation is logic – and logic is purely formal
    1. Purely formal processes can be run mechanically
2. The logic matters, not the architecture
    1. Implementation independence
        1. Humans
        2. Mechanics
        3. Electronics
3. Three things needed for computation
    1. Some input-output mechanisms (sensation)
    2. A big internal store (memory)
    3. Processes (programs) to do the logic

The computationalist claim

Human minds work the same way, i.e. Cognition is computation

- Computation is systematically interpretable symbol manipulation (it has strict syntax)
- Computation is implementation independent (it is multiply realizable, e.g. by mechanics, computers and brains)

What follows from the “cognition is computation” claim if *thinking* is part of human cognition?

How do we operationalise this?

The Turing Test YT Vid
.. if you cannot pass the test, it shows evidence for Turing indistinguishability (will be relevant for next week’s reading)

> !! Not just an operational definition of thinking, but of any cognitive process
> So..., if we can make our *computational cognitive model of, say, vision* work and break down in the same way as *human vision* works and breaks down, i.e. the two kinds of vision are *Turing indistinguishable*, then what follows?

Computationalism and psychology

Under behaviourism

- Psychology is a science of *behaviour*
- Our behaviour is determined by schedules of reinforcement

Under computationalism

- Psychology is a science of (human) computation
- Our psychology can be implemented in Turing Machines'

But how is cognitive science different from psychology?

Marr's three levels of analysis:

Computational theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Figure 1–4. The three levels at which any machine carrying out an information-processing task must be understood.

According to Marr:

"... the main job of vision [is] to derive a representation of shape"

Illusions and lesions may be especially revealing

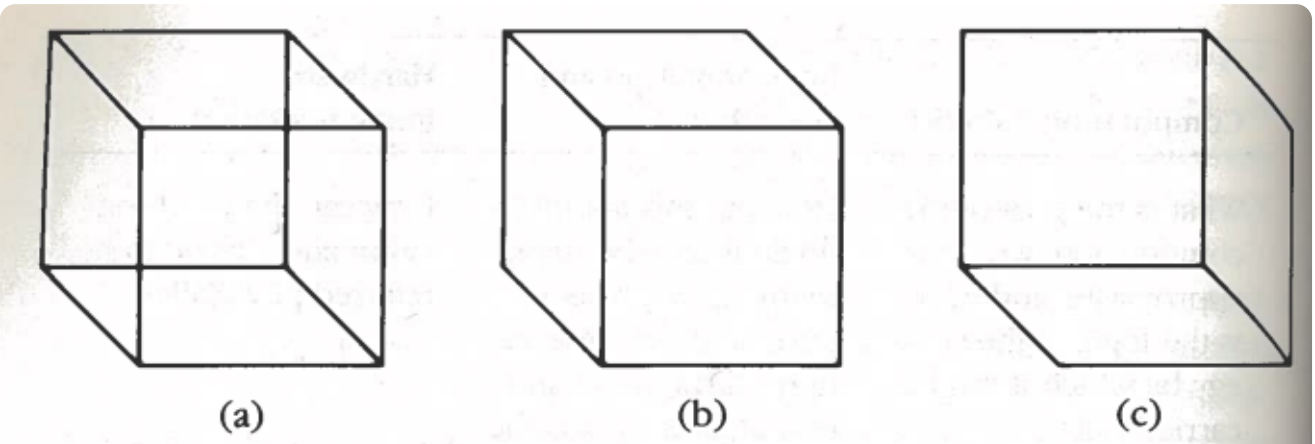
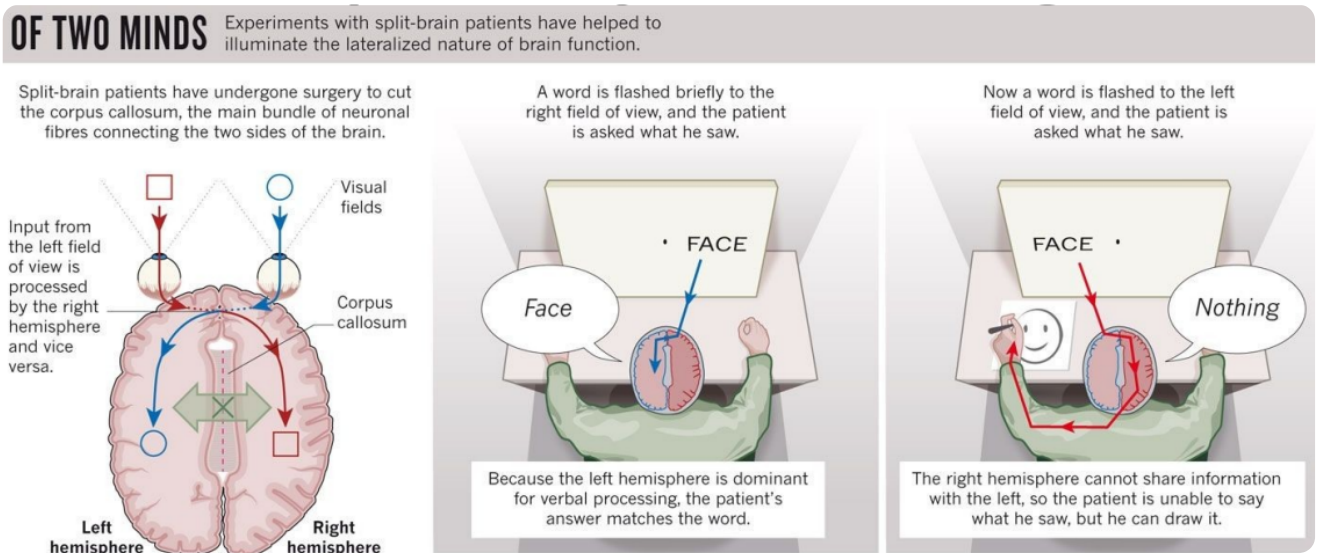


Figure 1–5. The so-called Necker illusion, named after L. A. Necker, the Swiss naturalist who developed it in 1832. The essence of the matter is that the two-dimensional representation (a) has collapsed the depth out of a cube and that a certain aspect of human vision is to recover this missing third dimension. The depth of the cube can indeed be perceived, but two interpretations are possible, (b) and (c). A person's perception characteristically flips from one to the other.

The computer model should "suffer" from the same bistability

experiments with split-brain patients have helped to illuminate the lateralized nature of brain function - this is actually really interesting stuff, which essentially implies that the

mind is two separate entities in need of a pathway to communicate



- The computer model should “suffer” from the same weakness (when emulating human vision)

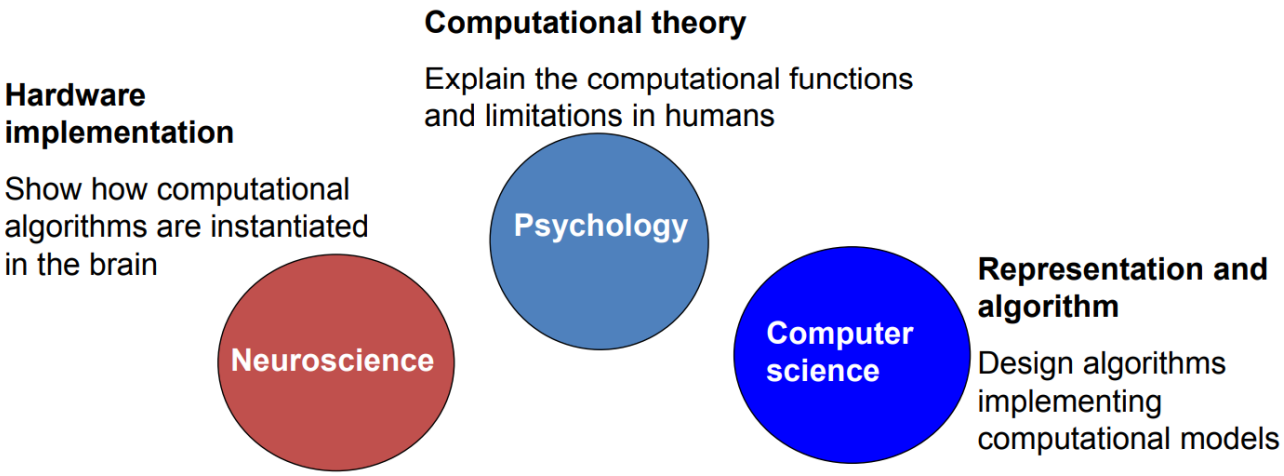
A quick aside: Paradigm shift:

- Something alien to behaviourism just appeared

What is the goal of the computation?

Marr's three levels (again)

- Use psychology to find the natural goal/function of the system investigated
- Use computer science to represent inputs and outputs, and to develop functions mapping inputs to outputs
- Do neuroscience to find out how the relevant algorithm is actually implemented (in the brain)



- This is the point where we have reached our *interdisciplinary* Cognitive Science

Rough translations to scientific practices

- Psychology = Experiments on function
- Computer Science = Modelling
- Neuroscience = Experiments on tissue

Interactions between the three levels

- Information from levels constrains formulation at adjacent levels
  - E.g. a computational theory may only permit certain representations
  - An implementation may only permit certain algorithms
- But they still shouldn't be blurred together
  - Each level of description constitutes its own epistemic domain

Now we have a programme!

> - do (cognitive) science according to the three levels and reveal facts about (human) cognition
> - Cognition is computation cognition is ⇒ implementation-independent, systematically interpretable symbol manipulation
> - ... **and a way to assess our models Turing** ⇒ indistinguishability

Marr - Vision, A Computational Investigation

Background

Early in this century, Wertheimer (1912, 1923) noticed the apparent motion not of individual dots but of wholes, or "fields", in images presented sequentially as in a movie. In much the say way we perceive the migration across the sky of a flock of geese: the flock somehow constitutes a single entity, and is not seen as individual birds.

- This observation started the Gestalt school of psychology (see Gestalt principles & Perception) which was concerned with describing the qualities of wholes by using terms like *solidarity* and *distinctness*, and with trying to formulate the "laws" that governed the creation of these wholes.
  - The attempt failed for various reasons, and the Gestalt school dissolved into the fog of subjectivism(?)

- With the death of the school, many of its early and genuine insights were unfortunately lost to the mainstream of experimental psychology.

(...) we know that the analysis of stereoscopic information, like the analysis of motion, can proceed independently in the absence of other information.

- Such findings are of critical importance because they can help us to subdivide our study of perception into more specialized parts which be treated separately.
- I shall refer to these as independent modules of perception

Marr talks about the Shepard & Metzler discovery of mental rotation (rotation time being linearly correlated with angle)
The significance of this approach lies not so much in its results but in the questions it raised

- Until then the notion of a representation was not one that visual psychologists took seriously, see Epiphenomenon (s/o Dennett being wrong as fuck)

A series of developments in neuroscience made it possible to study specific steps in the visual pathway one of which was Barlow's (1953) study of ganglion cells in the frog retina.

- The cumulative effect of the discoveries resulted in the realization that each *single neuron can perform a much more complex and subtle task than had previously been thought*

> "This amounts to a revolution in our outlook. It is now quite inappropriate to regard unit activity as a noisy indication of more basic and reliable processes involved in mental operations: instead, we must regard single neurons as the prime movers of theses mechanisms. Thinking is brought about by neurons and we should not use phrases like "unit activity reflects, reveals, or monitors thought processes," because the activities of neurons, quite simply, are thought processes."

This aspect of thinking led Barlow to formulate the first and most important of his five dogmas:

> "A description of that activity of a single nerve cell which is transmitted to and influences other nerve cells and of a nerve cell's response to such influences from other cells, is a complete enough description for functional understanding of the nervous system. There is nothing else "looking at" or

> controlling this activity, which must therefore provide a basis for understanding how the brain controls behavior.

- Marr reels in a bit here though: "I shall return later on to more carefully examine the validity of this point of view, but for now let us enjoy it"

Initially this reductionist approach seemed very promising (e.g. Gross, Rocha-Miranda and Bender, 1972, who found "hand-detectors" in the inferotemporal cortex)

Eventually though the apparent speedy progress slowed to a halt and a lot of open ends in the field of perception were left in the 70s

- "Suppose for example, that one actually found the apocryphal grandmother cell[1]. Would that really tell us that much at all?"
  - It doesn't tell us why or even how such a thing may be constructed from the outputs of previously discovered cells

Functional Heading

Attempting to recreate human vision in machines proved very hard

- There's a multitude of things to account for and an image is a complicated jumble of stuff
- The research did lead to some interesting questions about human perception though, like what does one actually do when a complete line drawing has been abstracted from a scene?

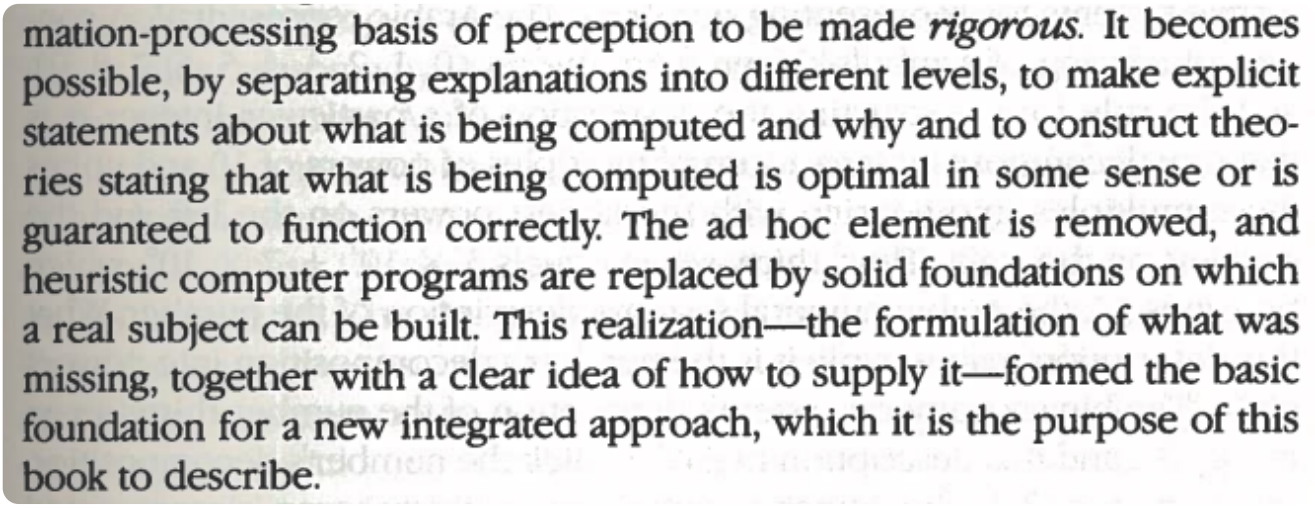
Research into these resulting questions brought about a clear message. There must exist and additional level of understanding at which the character of the information-processing tasks carried out during perception are analyzed and understood in a way that is independent of the particular mechanisms and structures that implement them in our heads.

- I.e. that vision and visual perception involves more systems than was known at the time

> !! "This was what was missing-- the analysis of the problem as information-processing task. Such analysis does not usurp an understanding at the other levels- of neurons or of computer programs- but it is a necessary

> complement to them, since without them there can be no real understanding of the function of all those neurons"

If the notion of different types of understanding is taken very seriously, it allows the study of the information-processing basis of perception to be made *rigorous*.



mation-processing basis of perception to be made *rigorous*. It becomes possible, by separating explanations into different levels, to make explicit statements about what is being computed and why and to construct theories stating that what is being computed is optimal in some sense or is guaranteed to function correctly. The ad hoc element is removed, and heuristic computer programs are replaced by solid foundations on which a real subject can be built. This realization—the formulation of what was missing, together with a clear idea of how to supply it—formed the basic foundation for a new integrated approach, which it is the purpose of this book to describe.

Understanding Complex Information-Processing Systems

> !! Almost never can a complex system of any kind be understood as a simple extrapolation from the properties of its elementary components

For the specific case of a system that solves an information-processing problem, there are in addition the twin strands of process and representation, and both these ideas need some discussion.

Representation and Description

A *representation* is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this. I shall call the result of using a representation to describe a given entity a *description* of the entity in that representation (Marr and Nishihara, 1978)

For example, the Arabic, Roman and binary numeral systems are all formal systems for representing numbers.

The Arabic *representation* consists of a string of symbols drawn from the set (0,1,2,3,4,5,6,7,8,9), and the rule for constructing the description of a particular integer n is that one decomposes n into a sum of multiples of powers of 10 and unites these into a string with the largest powers on the left and the smallest on the right.

- e.g. thirty-seven equals $3 \times 10^1 + 7 \times 10^0$ (the Arabic numeral systems *description* of the number)
    - What this description makes explicit is the number's decomposition into powers of 10

The binary numeral system's description of the number thirty seven is 100101, and this description makes explicit the number's decomposition into powers of 20.

In the Roman numeral system, thirty-seven is represented as $\mathbf{XXXVII}$

This definition of a representation is quit general. A musical score provides a way of representing a symphony; the alphabet allows the construction of a written representation of words. The phrase "formal scheme" is critical to the definition.

Formal scheme: a set of symbols with rules for putting them together

Any particular representation makes certain information explicit at the expense of information that is pushed into the background and may be quite hard to recover.

Process

The term *process* is very broad. For the purposes of this book, I want to restrict our attention to the meanings associated with machines that are carrying out information-processing tasks.

There are several layers and it is perhaps most useful to think in terms of three

The most abstract level is the level of *what* the device does and *why*.
In the case of a cash register, what it does is arithmetic.

- In this case then it has the task of mastering addition
- There are fundamental properties which are part of the fundamental *theory* of addition (commutative, associative etc.), which apply to all number systems (they are true in essence)
- Thus part of this first level is something that might be characterized as *what* is being computed

The other half of this level is the *why* the register performs addition and not for instance multiplication. The reason is that the rules we intuitively feel to be appropriate for combining the individual prices in fact define the mathematical operation of addition. These can be formulated as *constraints* in the following way: (basically)

1. Buying nothing should cost nothing (rule for zero)
2. Order should not affect the total (commutativity)
3. Paying for piles separately should not affect total (associativity)
4. Buying an item and returning should total to zero (inverses)

It is a mathematical theorem that these conditions define the operation of addition, which is therefore the appropriate computation to use

> "This whole argument is what I call the *computational theory* of the cash register"

Its important features are (1) that it contains separate arguments about what is computed and why and (2) that the resulting operation is defined uniquely by the constraints it has to satisfy.

In the theory of visual processes, the underlying task is to reliably derive properties of the world from images of it; the business of isolating constraints that are both powerful enough to allow a process to be defined and generally true of the world is a central theme of our inquiry.

In order to actually run a process, however, one has to realize it in some way and therefore choose a representation for the entities that the process manipulates.

The second level is therefore involves choosing two things: (1) a *representation* for the input and for the output of the process and (2) an *algorithm* by which the transformation may actually be accomplished.

- If the first of our levels specify what and why, this second level specifies *how*

There are usually a wide array of choice of representation, which the algorithm often rather critically depends on. For a given fixed representation there are often multiple ways of carrying out algorithms for the same process.

- One algorithm may be more efficient although less robust and vice versa
- One may be serial and another parallel and the choice may depend on what hardware the algorithm has to be run on

The third level is that of the device in which the process is to be realized physically. The important point here is that, once again, the same algorithm may be implemented in quite different technologies.

The Three Levels

> "Trying to understand perception by studying only neurons is like trying to understand flight by studying feathers: it just can't be done" - David Marr

Computational theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Figure 1–4. The three levels at which any machine carrying out an information-processing task must be understood.

There is a wide choice available at each level and the explication of each level involves issues that are rather independent of the other two

An important point to note is that since the three levels are only rather loosely related, some phenomena may be explained at only one or two of them.

- In psychophysiology some things are related to the physical mechanisms of vision (afterimages of a lightbulb for instance)
- On the other hand the ambiguity of the Necker Cube (see below) seems to demand a different explanation

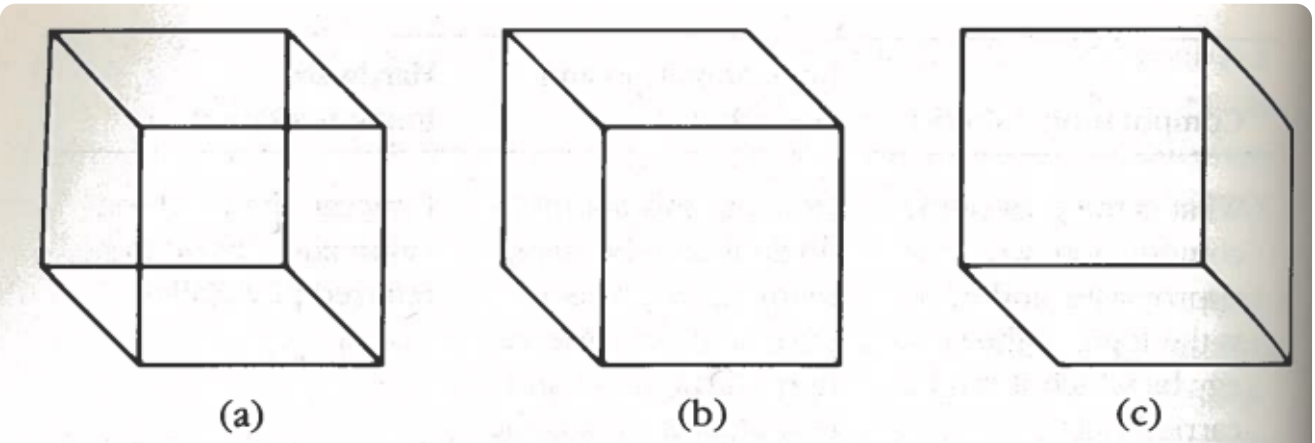


Figure 1–5. The so-called Necker illusion, named after L. A. Necker, the Swiss naturalist who developed it in 1832. The essence of the matter is that the two-dimensional representation (a) has collapsed the depth out of a cube and that a certain aspect of human vision is to recover this missing third dimension. The depth of the cube can indeed be perceived, but two interpretations are possible, (b) and (c). A person's perception characteristically flips from one to the other.

One objection raised to computationalism is that the brain is quite different from a computer because one is parallel and one is serial. The answer is that the distinction between serial and parallel is a distinction at the level of algorithm; it is not fundamental. Anything programmed in in parallel can be rewritten serially (not necessarily vice versa tho)

Importance of Computational Theory

Although algorithms and mechanisms are empirically more accessible, it is the top level, the level of computational theory, which is critically important from an information-processing point of view.

- An algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied

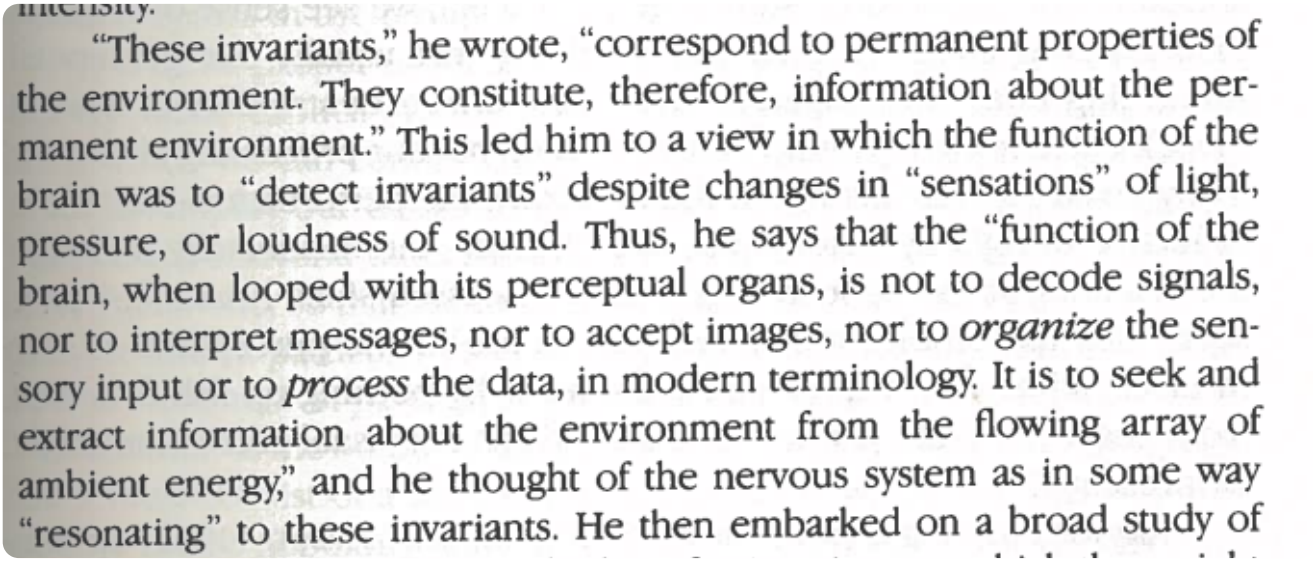
Regarding AI, for far too long a heuristic program for carrying out some task was held to be a theory of that task and the distinction between *what* a program did and *how* it did it was not taken seriously

Chomsky's (1965) theory of transformational grammar is a true computational theory in the sense defined earlier. It is concerned solely with specifying what the syntactic decomposition of an English sentences should be, and not at all with how that decomposition should be achieved.

The Approach of J. J. Gibson

Gibson (1966) was perhaps the one who came closest to a perceptional computational theory. He asked the right question: How does one obtain constant perceptions in everyday life on the basis of continually changing sensations?

- He oversimplified his view of how such a model should be done however



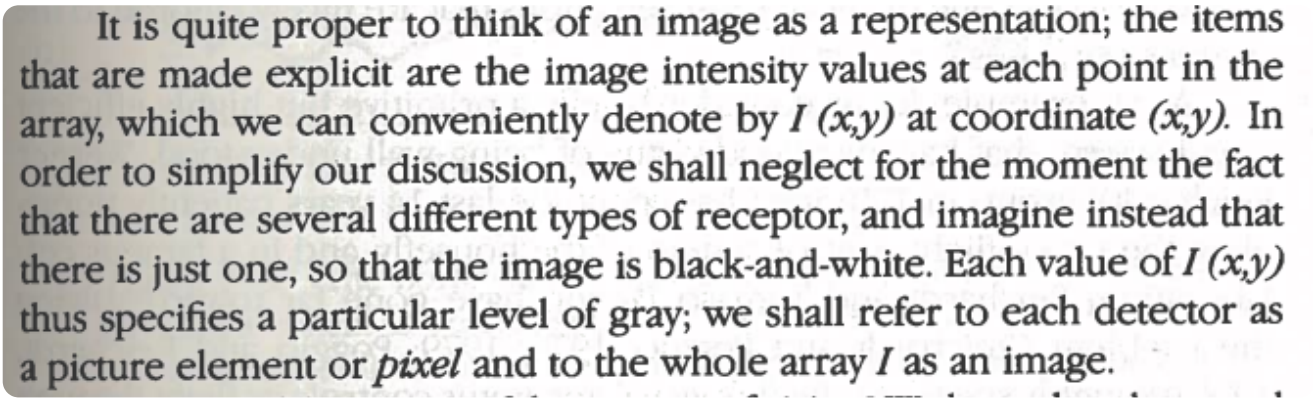
intensity.

"These invariants," he wrote, "correspond to permanent properties of the environment. They constitute, therefore, information about the permanent environment." This led him to a view in which the function of the brain was to "detect invariants" despite changes in "sensations" of light, pressure, or loudness of sound. Thus, he says that the "function of the brain, when looped with its perceptual organs, is not to decode signals, nor to interpret messages, nor to accept images, nor to *organize* the sensory input or to *process* the data, in modern terminology. It is to seek and extract information about the environment from the flowing array of ambient energy," and he thought of the nervous system as in some way "resonating" to these invariants. He then embarked on a broad study of

It's apparently quite common in the tradition of inquiry into perception to underestimate its complexity.

A Representational Framework For Vision

Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information (Marr, 1976)



It is quite proper to think of an image as a representation; the items that are made explicit are the image intensity values at each point in the array, which we can conveniently denote by $I(x,y)$ at coordinate (x,y). In order to simplify our discussion, we shall neglect for the moment the fact that there are several different types of receptor, and imagine instead that there is just one, so that the image is black-and-white. Each value of $I(x,y)$ thus specifies a particular level of gray; we shall refer to each detector as a picture element or *pixel* and to the whole array I as an image.

The Purpose of Vision

Vision, in short, is used in such a bewildering variety of ways that the visual systems of different animals must differ significantly from one another.

In-flight control is achieved by independent systems controlling the fly's vertical velocity (through control of the lift generated by the wings) and horizontal direction (determined by the torque produced by the asymmetry of the horizontal thrust from the left and right wings). The visual input to the horizontal control system, for example, is completely described by the two terms

$$r(\psi)\dot{\psi} + D(\psi)$$

where r and D have the form illustrated in Figure 1–6. This input describes how the fly tracks an object that is present at angle ψ in the visual field and has angular velocity $\dot{\psi}$. This system is triggered to track objects of a certain angular dimension in the visual field, and the motor strategy is such that if the visible object was another fly a few inches away, then it would be

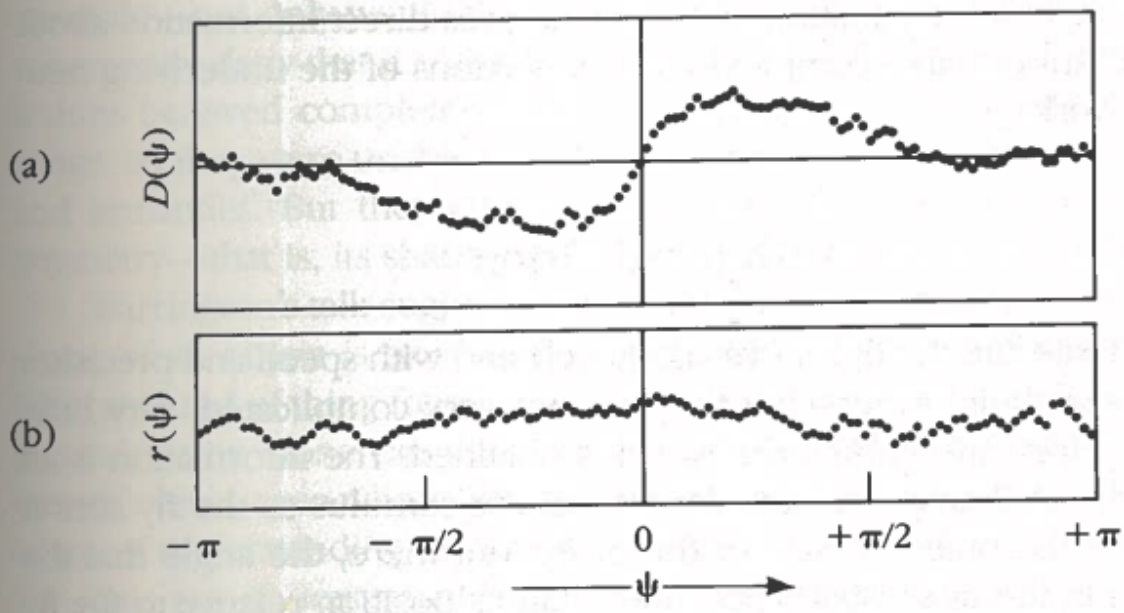


Figure 1–6. The horizontal component of the visual input R to the fly's flight system is described by the formula $R = D(\psi) - r(\psi)\dot{\psi}$, where ψ is the direction of the stimulus and $\dot{\psi}$ is its angular velocity in the fly's visual field. $D(\psi)$ is an odd function, as shown in (a), which has the effect of keeping the target centered in the fly's visual field; $r(\psi)$ is essentially constant as shown in (b).

In particular it is extremely unlikely that the fly has any explicit representation of the visual world around him-- no true conception of a surface, for example, but just a few triggers and some specifically fly-centered parameters like ψ and $\dot{\psi}$.

Human vision is more complex although not unlike the fly's such that it may well incorporate subsystems to help with specific low-level tasks like the control of pursuit eye movements.

Advanced Vision

Warrington and Taylor (1973) studied capacities of patients with parietal lesions (L/R) Warrington's talk suggested two things

- The representation of an object is stored a different place and is therefore quite different from the representation of its use and purpose
- Vision alone can deliver an internal description of the shape of a viewed object, even when the object was not recognized in the conventional sense of understanding its use and purpose

Warrington had put her finger on what was the quintessential fact of human vision--that it tells about shape and space and spatial arrangement. It does other things of course, but these could be hung off a theory in which the main job of vision is to derive a representation of shape.

To the Desirable via the Possible

... the main stepping stone toward this goal is describing the geometry of the visible surfaces, shading, texture, contours, or visual motion, is due to a shape's local surface properties. The objective of many early visual computations is to extract this information

However this description of the visible surfaces turns out to be unsuitable for recognition tasks. There are several reasons why, perhaps the most prominent being that all early visual processes, it depends critically on perspective (vantage point)

The final step therefore consists of transforming the viewer-centered surface description into a representation of the three-dimensional shape and spatial arrangement of an object that does not depend upon the direction from which the object is being viewed.

- This final description is object centered rather than viewer centered.

Table 1–1. Representational framework for deriving shape information from images.

Name	Purpose	Primitives
Image(s)	Represents intensity.	Intensity value at each point in the image
Primal sketch	Makes explicit important information about the two-dimensional image, primarily the intensity changes there and their geometrical distribution and organization.	Zero-crossings Blobs Terminations and discontinuities Edge segments Virtual lines Groups Curvilinear organization Boundaries
2½-D sketch	Makes explicit the orientation and rough depth of the visible surfaces, and contours of discontinuities in these quantities in a viewer-centered coordinate frame.	Local surface orientation (the "needles" primitives) Distance from viewer Discontinuities in depth Discontinuities in surface orientation
3-D model representation	Describes shapes and their spatial organization in an object-centered coordinate frame, using a modular hierarchical representation that includes volumetric primitives (i.e., primitives that represent the volume of space that a shape occupies) as well as surface primitives.	3-D models arranged hierarchically, each one based on a spatial configuration of a few sticks or axes, to which volumetric or surface shape primitives are attached

The overall framework described here therefore divides the derivation of shape information from images into three representational stages (seen above)

1. The representation of properties of the two-dimensional image, such as intensity changes and local two-dimensional geometry
2. the representation of properties of the visible surfaces in a viewer-centered coordinate system, such as surface orientation, distance from the viewer, and discontinuities in these quantities; surface reflectance; and some coarse description of the prevailing illumination
3. an object centered representation of the three-dimensional structure and of the organization of the viewed shape together with some description of its surface properties.

Reading Questions

Marr:

1. What does Marr mean when he says that "understanding computers is different from understanding computations" (p. 5)?
    1. Marr argues that understanding computers involves understanding their hardware (i.e. their physical components) and how they operate whereas understanding computations involves understanding what they're doing, i.e. what are the algorithms doing (logical processes)
2. Why do you think Marr chose vision as his preferred example for demonstrating how the mind works? (P.S. There could be multiple good answers.)
    1. It's the most dominant modality we posses, and a lot of early work in psychophysiology centered around vision and the neurons responsible for parts of it
    2. Vision is also a deeply complex process that illustrates how the mind makes sense of the world
3. Why were the studies on receptive fields (e.g. Hubel & Wiesel) and mental rotation (e.g., Shepard & Metzler), respectively, so instrumental in paving the way for computationalism?
    1. "The significance of this approach lies not so much in its results but in the questions it raised. Until then the notion of a representation was not one that visual psychologists took seriously
    2. Both also provided insights into the neural mechanisms underlying visual processing, which were essential for developing computational models of vision
4. Why did scientists leave behind the reductionist approach (described on pp. 12-14)?
    1. After the initial discoveries made in the 50s and 60s, progress halted and it became apparent that more layers had to be involved than just the neurons in themselves being responsible for the mind's activity
    2. Marr argues that a multi-level approach is essential to understanding cognitive processes (vision/perception)
5. What is Marr's main argument why we need multiple levels of explanation? Is that a good argument in your view?
    1. We need to be able to do more than *describe* we have to be capable of *explaining* as well. It's an alright argument
6. How exactly does Marr define "representations"?

    1. A *representation* is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this
        1. Marr gives numeral systems as example (quantities having different representations of the same thing essentially)
7. Why do you think representations and algorithms both inhabit the same conceptual level and not two separate levels in Marr's model?
    1. Because they're inseparable entities relying on each other. You need one to understand the other
        1. Representations serve as the input and output of algorithms, and the choice of algorithm depends on the nature of the representations and the task requirements.
8. Which criteria typically go into picking an appropriate algorithm for a given task?
    1. What the specific needs are for the task and what hardware the algorithm has to run on
    2. Criteria such as efficiency, accuracy, robustness, and computational complexity are typically considered.
9. Marr admits that "some phenomena may be explained at only one or two [levels]" (p. 25). What about explaining phenomena at more than three levels? Can you think of any cases where that might be useful or necessary?
    1. While Marr suggests that most phenomena can be explained at multiple levels, there may be cases where explaining phenomena at more than three levels could be useful or necessary.
    2. For example, complex cognitive processes involving interactions between multiple brain regions may require explanations at both neural and algorithmic levels.
10. According to Marr, there is an asymmetry between the importance of the three levels. What is the nature of this asymmetry, and why is it there?
    1. Marr proposes an asymmetry in the importance of the three levels of analysis, where the computational level is considered the most important, followed by the algorithmic level and then the implementation level.
    2. This hierarchy reflects the idea that understanding the computational goals and strategies is fundamental to understanding cognitive processes.
11. Does vision in different animals have: (a) the same purpose and same representations; (b) the same purpose but different representations; (c) different purposes but same representations; or (d) different purposes and different representations?
    1. Different purposes and different representations I would imagine

    2. Vision in different animals may have different purposes and representations. For example, while humans and certain animals share the basic purpose of detecting and recognizing objects in their environment, the specific representations and processing mechanisms may differ based on factors such as ecological niche and evolutionary history.

12. Is the process of recognizing objects "viewer-centred" or "object-centred"? In your own words, what does this mean?
    1. It begins viewer-centered but has as a goal to be object-centered
    2. It means that for us to recognize objects we must recognize them independently of viewing angle and vantage point (I believe)
    3. The process of recognizing objects can be described as either viewer-centered or object-centered. Viewer-centered recognition refers to recognizing objects relative to the observer's viewpoint or perspective, while object-centered recognition involves identifying objects based on their intrinsic properties, independent of the observer's viewpoint. This distinction highlights different strategies used by the cognitive system to interpret visual stimuli.

Footnotes

1. A cell that only fires when one's grandmother comes into view (an example of the reductionist view that there exists cells for every recognizable entity)↩

The Computational Theory of Mind

- The text from the Stanford Encyclopedia for next time is quite comprehensive. Focus on sections 1,2, 3 and 5 if you get lost

Computational Mind

Advances in computing raise the prospect that the mind itself is a computational system—a position known as *the computational theory of mind* (CTM). *Computationalists* are researchers who endorse CTM, at least as applied to certain important mental processes. CTM played a central role within cognitive science during the 1960s and 1970s. For many years, it enjoyed orthodox status. More recently, it has come under pressure from various rival paradigms.

- A key task facing computationalists is to explain what one means when one says that the mind “computes”.
- A second task is to argue that the mind “computes” in the relevant sense.
- A third task is to elucidate how computational description relates to other common types of description, especially *neurophysiological description* (which cites neurophysiological properties of the organism’s brain or body) and *intentional description* (which cites representational properties of mental states).

Turing Machines

The intuitive notions of *computation* and *algorithm* are central to mathematics. Roughly speaking, an algorithm is an explicit, step-by-step procedure for answering some question or solving some problem. An algorithm provides *routine mechanical instructions* dictating how to proceed at each step. See Marr - Vision, A Computational Investigation > Process

Alan Turing’s landmark paper “On Computable Numbers, With an Application to the Entscheidungsproblem” (Turing 1936) offered the analysis that has proved most influential.

Turing motivates his approach by reflecting on idealized human computing agents. Citing finitary limits on our perceptual and cognitive apparatus, he argues that any symbolic algorithm executed by a human can be replicated by a suitable Turing machine. He concludes that the Turing machine formalism, despite its extreme simplicity, is powerful enough to capture all humanly executable mechanical procedures over symbolic configurations. Subsequent discussants have almost universally agreed.

Turing computation is often described as *digital* rather than *analog*. What this means is not always so clear, but the basic idea is usually that computation operates over discrete configurations. By comparison, many historically important algorithms operate over continuously variable configurations. For example, Euclidean geometry...

- Turing machines operate over discrete strings of elements (digits) drawn from a finite alphabet.
- One recurring controversy concerns whether the digital paradigm is well-suited to model mental activity or whether an analog paradigm would instead be more fitting (MacLennan 2012; Piccinini and Bahar 2013).

Besides introducing Turing machines, Turing (1936) proved several seminal mathematical results involving them. In particular, he proved the existence of a *universal Turing machine* (UTM).

Roughly speaking, a UTM is a Turing machine that can mimic any other Turing machine. One provides the UTM with a symbolic input that codes the machine table for Turing machine *M*. The UTM replicates *M*'s behavior, executing instructions enshrined by *M*'s machine table. In that sense, the UTM is a *programmable general purpose computer*.

- To a first approximation, all personal computers are also general purpose: they can mimic any Turing machine, when suitably programmed.
- The main caveat is that physical computers have finite memory, whereas a Turing machine has unlimited memory.
    - More accurately, then, a personal computer can mimic any Turing machine *until it exhausts its limited memory supply*.

Turing's discussion helped lay the foundations for *computer science*, which seeks to design, build, and understand computing systems. As we know, computer scientists can now build extremely sophisticated computing machines. All these machines implement something resembling Turing computation, although the details differ from Turing's simplified model

Artificial Intelligence

Rapid progress in computer science prompted many, including Turing, to contemplate whether we could build a computer capable of thought. *Artificial Intelligence* (AI) aims to construct "thinking machinery".
More precisely, it aims to construct computing machines that execute core mental tasks such as reasoning, decision-making, problem solving, and so on.

- During the 1950s and 1960s, this goal came to seem increasingly realistic (Haugeland 1985).

Early AI research emphasized *logic*. Researchers sought to "mechanize" deductive reasoning. A famous example was the *Logic Theorist* computer program (Newell and Simon 1956), which proved 38 of the first 52 theorems from *Principia Mathematica* (Whitehead and Russell 1925). In one case, it discovered a simpler proof than *Principia*'s.

When confident predictions of thinking machines proved too optimistic, many observers lost interest or concluded that AI was a fool's errand. Nevertheless, the decades have witnessed gradual progress.

- One striking success was IBM's Deep Blue, which defeated chess champion Gary Kasparov in 1997.
- Another major success was the driverless car Stanley (Thrun, Montemerlo, Dahlkamp, et al. 2006), which completed a 132-mile course in the Mojave Desert, winning the 2005 Defense Advanced Research Projects Agency (DARPA) Grand Challenge.
- A less flashy success story is the vast improvement in speech recognition algorithms.

One problem that dogged early work in AI is *uncertainty*. Nearly all reasoning and decision-making operates under conditions of uncertainty. For example, you may need to decide whether to go on a picnic while being uncertain whether it will rain. *Bayesian decision theory* is the standard mathematical model of inference and decision-making under uncertainty. Uncertainty is codified through *probability*. Precise rules dictate how to update probabilities in light of new evidence and how to select actions in light of probabilities and utilities.

- See The Brain and Decision Making & Decision Making Class Notes

In the 1980s and 1990s, technological and conceptual developments enabled efficient computer programs that implement or approximate Bayesian inference in realistic scenarios.

- An explosion of Bayesian AI ensued (Thrun, Burgard, and Fox 2006), including the aforementioned advances in speech recognition and driverless vehicles.
- Tractable algorithms that handle uncertainty are a major achievement of contemporary AI (Murphy 2012), and possibly a harbinger of more impressive future progress.

Some philosophers insist that computers, no matter how sophisticated they become, will at best *mimic* rather than *replicate* thought

Turing (1950) anticipated these worries and tried to defuse them. He proposed a scenario, now called *the Turing Test*, where one evaluates whether an unseen interlocutor is a computer or a human.

- A computer *passes the Turing test* if one cannot determine that it is a computer. Turing proposed that we abandon the question "Could a computer think?" as

hopelessly vague, replacing it with the question “Could a computer pass the Turing test?”.

- Turing’s discussion has received considerable attention, proving especially influential within AI.
  - Ned Block (1981) offers an influential critique. He argues that certain possible machines pass the Turing test even though these machines do not come close to genuine thought or intelligence.

The Classical Computational Theory of Mind

Warren McCulloch and Walter Pitts (1943) first suggested that something resembling the Turing machine might provide a good model for the mind.

> In the 1960s, Turing computation became central to the emerging interdisciplinary initiative *cognitive science*, which studies the mind by drawing upon psychology, computer science (especially AI), linguistics, philosophy, economics (especially game theory and behavioral economics), anthropology, and neuroscience

The label *classical computational theory of mind* (which we will abbreviate as CCTM) is now fairly standard. According to CCTM, the mind is a computational system similar in important respects to a Turing machine, and core mental processes (e.g., reasoning, decision making, and problem solving) are computations similar in important respects to computations executed by a Turing machine. These formulations are imprecise. CCTM is best seen as a family of views, rather than a single well-defined view.[1]

It is common to describe CCTM as embodying “the computer metaphor”. This description is doubly misleading.

First, CCTM is better formulated by describing the mind as a “computing system” or a “computational system” rather than a “computer”.

- As David Chalmers (2011) notes, describing a system as a “computer” strongly suggests that the system is *programmable*. (...) Critics of CCTM often object that the mind is not a programmable general purpose computer (Churchland, Koch, and Sejnowski 1990).
- Since classical computationalists need not claim (and usually do not claim) that the mind is a programmable general purpose computer, the objection is misdirected.

Second, CCTM is not intended metaphorically. CCTM does not simply hold that the mind is *like* a computing system. CCTM holds that the mind *literally is* a computing system. Of course, the most familiar artificial computing systems are made from silicon chips or similar materials, whereas the human body is made from flesh and blood. But CCTM holds that this difference disguises a more fundamental similarity, which we can capture through a Turing-style computational model. In offering such a model, we prescind from physical details.

- We attain an abstract computational description that could be physically implemented in diverse ways (e.g., through silicon chips, or neurons, or pulleys and levers).
- CCTM holds that a suitable abstract computational model offers a literally true description of core mental processes.

It is common to summarize CCTM through the slogan “the mind is a Turing machine”. This slogan is also somewhat misleading, because no one regards Turing’s precise formalism as a plausible model of mental activity. The formalism seems too restrictive in several ways:

1. Turing machines execute pure symbolic computation. The inputs and outputs are symbols inscribed in memory locations. In contrast, the mind receives *sensory input* (e.g., retinal stimulations) and produces *motor output* (e.g., muscle activations). A complete theory must describe how mental computation interfaces with sensory inputs and motor outputs.
2. A Turing machine has infinite discrete memory capacity. Ordinary biological systems have finite memory capacity. A plausible psychological model must replace the infinite memory store with a large but finite memory store
3. Modern computers have *random access memory* (s/o Daft Punk): addressable memory locations that the central processor can directly access. Turing machine memory is not addressable. The central processor can access a location only by sequentially accessing intermediate locations.
    1. Computation without addressable memory is hopelessly inefficient. For that reason, C.R. Gallistel and Adam King (2009) argue that addressable memory gives a better model of the mind than non-addressable memory.
4. A Turing machine has a central processor that operates *serially*, executing one instruction at a time. Other computational formalisms relax this assumption, allowing multiple processing units that operate in *parallel*. Classical computationalists can allow parallel computations (Fodor and Pylyshyn 1988; Gallistel and King 2009: 174)
5. Turing computation is *deterministic*: total computational state determines subsequent computational state. One might instead allow *stochastic* computations.

In a stochastic model, current state does not dictate a unique next state. Rather, there is a certain probability that the machine will transition from one state to another.

CCTM claims that mental activity is “Turing-style computation”, allowing these and other departures from Turing’s own formalism.

Machine Functionalism

Hilary Putnam (1967) introduced CCTM into philosophy. He contrasted his position with *logical behaviorism* and *type-identity theory*. Each position purports to reveal the nature of mental states, including propositional attitudes (e.g., beliefs), sensations (e.g., pains), and emotions (e.g., fear).

- According to logical behaviorism, mental states are behavioral dispositions.
- According to type-identity theory, mental states are brain states.

Putnam advances an opposing *functionalist* view, on which mental states are functional states.

- According to functionalism, a system has a mind when the system has a suitable *functional organization*.
- Mental states are states that play appropriate roles in the system’s functional organization. Each mental state is individuated by its interactions with sensory input, motor output, and other mental states.

Functionalism offers notable advantages over logical behaviorism and type-identity theory:

- Behaviorists want to associate each mental state with a characteristic pattern of behavior—a hopeless task, because individual mental states do not usually have characteristic behavioral effects.
  - Behavior almost always results from distinct mental states operating together (e.g., a belief and a desire).
  - Functionalism avoids this difficulty by individuating mental states through characteristic relations not only to sensory input and behavior but also to one another.
- Type-identity theorists want to associate each mental state with a characteristic physical or neurophysiological state.

- Putnam casts this project into doubt by arguing that mental states are *multiply realizable*: the same mental state can be realized by diverse physical systems, including not only terrestrial creatures but also hypothetical creatures (e.g., a silicon-based Martian).
- Functionalism is tailor-made to accommodate multiple realizability. According to functionalism, what matters for mentality is a pattern of organization, which could be physically realized in many different ways.

Putnam defends a brand of functionalism now called *machine functionalism*. He emphasizes *probabilistic automata*, which are similar to Turing machines except that transitions between computational states are stochastic.

- He proposes that mental activity implements a probabilistic automaton and that particular mental states are machine states of the automaton's central processor.
- The machine table specifies an appropriate functional organization, and it also specifies the role that individual mental states play within that functional organization. In this way, Putnam combines functionalism with CCTM.

Machine functionalism faces several problems.
One problem, highlighted by Ned Block and Jerry Fodor (1972), concerns the *productivity of thought*. A normal human can entertain a potential infinity of propositions. Machine functionalism identifies mental states with machine states of a probabilistic automaton.

- Since there are only finitely many machine states, there are not enough machine states to pair one-one with possible mental states of a normal human. Of course, an actual human will only ever entertain finitely many propositions.
- However, Block and Fodor contend that this limitation reflects limits on lifespan and memory, rather than (say) some psychological law that restricts the class of humanly entertainable propositions.
- A probabilistic automaton is endowed with unlimited time and memory capacity yet even still has only finitely many machine states. Apparently, then, machine functionalism mislocates the finitary limits upon human cognition.

Another problem for machine functionalism, also highlighted by Block and Fodor (1972), concerns the *systematicity of thought.* An ability to entertain one proposition is correlated with an ability to think other propositions. For example, someone who can entertain the thought *that John loves Mary* can also entertain the thought *that Mary loves John*. Thus, there seem to be systematic relations between mental states.

- A good theory should reflect those systematic relations. Yet machine functionalism identifies mental states with unstructured machines states, which lack the

requisite systematic relations to another.

- For that reason, machine functionalism does not explain systematicity.
- In response to this objection, machine functionalists might deny that they are obligated to explain systematicity.
- Nevertheless, the objection suggests that machine functionalism neglects essential features of human mentality. A better theory would explain those features in a principled way.

While the productivity and systematicity objections to machine functionalism are perhaps not decisive, they provide strong impetus to pursue an improved version of CCTM.

The Representational Theory of Mind

> Fodor (1975, 1981, 1987, 1990, 1994, 2008) advocates a version of CCTM that accommodates systematicity and productivity much more satisfactorily. He shifts attention to the *symbols* manipulated during Turing-style computation.

An old view, stretching back at least to William of Ockham's *Summa Logicae*, holds that thinking occurs in a *language of thought* (sometimes called *Mentalese*). Fodor revives this view. He postulates a system of mental representations, including both primitive representations and complex representations formed from primitive representations.

- See The Language of Thought Hypothesis

For example, the primitive Mentalese words JOHN, MARY, and LOVES can combine to form the Mentalese sentence JOHN LOVES MARY. Mentalese is *compositional*: the meaning of a complex Mentalese expression is a function of the meanings of its parts and the way those parts are combined.

- Propositional attitudes are relations to Mentalese symbols.
- Fodor calls this view *the representational theory of mind* (*RTM*).
- Combining RTM with CCTM, he argues that mental activity involves Turing-style computation over the language of thought.
- Mental computation stores Mentalese symbols in memory locations, manipulating those symbols in accord with mechanical rules.

A prime virtue of RTM is how readily it accommodates productivity and systematicity

Productivity: RTM postulates a finite set of primitive Mentalese expressions, combinable into a potential infinity of complex Mentalese expressions. A thinker with access to primitive Mentalese vocabulary and Mentalese compounding devices has the potential to entertain an infinity of Mentalese expressions. She therefore has the potential to instantiate infinitely many propositional attitudes (neglecting limits on time and memory).

Systematicity: According to RTM, there are systematic relations between which propositional attitudes a thinker can entertain. For example, suppose I can think that John loves Mary. According to RTM, my doing so involves my standing in some relation *R* to a Mentalese sentence JOHN LOVES MARY, composed of Mentalese words JOHN, LOVES, and MARY combined in the right way. If I have this capacity, then I also have the capacity to stand in relation *R* to the distinct Mentalese sentence MARY LOVES JOHN, thereby thinking that Mary loves John. So the capacity to think that John loves Mary is systematically related to the capacity to think that Mary loves John.

By treating propositional attitudes as relations to complex mental symbols, RTM explains both productivity and systematicity.

CCTM+RTM differs from machine functionalism in several other respects.

- First, machine functionalism is a theory of mental states *in general*, while RTM is only a theory of propositional attitudes.
- Second, proponents of CCTM+RTM need not say that propositional attitudes are individuated functionally.
- As Fodor (2000: 105, fn. 4) notes, we must distinguish *computationalism* (mental processes are computational) from *functionalism* (mental states are functional states). Machine functionalism endorses both doctrines. CCTM+RTM endorses only the first.

Philosophical discussion of RTM tends to focus mainly on *high-level human thought*, especially belief and desire. However, CCTM+RTM is applicable to a much wider range of mental states and processes. Many cognitive scientists apply it to non-human animals. For example, Gallistel and King (2009) apply it to certain invertebrate phenomena (e.g., honeybee navigation).
Even confining attention to humans, one can apply CCTM+RTM to *subpersonal processing*. Fodor (1983) argues that perception involves a subpersonal “module” that converts retinal input into Mentalese symbols and then performs computations over those symbols.

- Thus, talk about a language of *thought* is potentially misleading, since it suggests a non-existent restriction to higher-level mental activity.

Also potentially misleading is the description of Mentalese as a *language*, which suggests that all Mentalese symbols resemble expressions in a natural language. Many philosophers, including Fodor, sometimes seem to endorse that position. However, there are possible non-propositional formats for Mentalese symbols. Proponents of CCTM+RTM can adopt a pluralistic line, allowing mental computation to operate over items akin to images, maps, diagrams, or other non-propositional representations[2]

The pluralistic line seems especially plausible as applied to subpersonal processes (such as perception) and non-human animals. Michael Rescorla (2009a,b) surveys research on *cognitive maps* (see Cognitive Maps) (Tolman 1948; O'Keefe and Nadel 1978; Gallistel 1990), suggesting that some animals may navigate by computing over mental representations more similar to maps than sentences.
Elisabeth Camp (2009), citing research on baboon social interaction (Cheney and Seyfarth 2007), argues that baboons may encode social dominance relations through non-sentential (not relating to a sentence, red) tree-structured representations.

CCTM+RTM is *schematic*. To fill in the schema, one must provide detailed computational models of specific mental processes. A complete model will:

- describe the Mentalese symbols manipulated by the process;
- isolate elementary operations that manipulate the symbols (e.g., *inscribing a symbol in a memory location*); and
- delineate mechanical rules governing application of elementary operations.

By providing a detailed computational model, we decompose a complex mental process into a series of elementary operations governed by precise, routine instructions.

CCTM+RTM remains neutral in the traditional debate between physicalism and substance dualism. A Turing-style model proceeds at a very abstract level, not saying whether mental computations are implemented by physical stuff or Cartesian soul-stuff (Block 1983: 522). In practice, all proponents of CCTM+RTM embrace a broadly physicalist outlook.

- They hold that mental computations are implemented not by soul-stuff but rather by the brain.
- On this view, Mentalese symbols are realized by neural states, and computational operations over Mentalese symbols are realized by neural processes.

Ultimately, physicalist proponents of CCTM+RTM must produce empirically well-confirmed theories that explain how exactly neural activity implements Turing-style computation. As Gallistel and King (2009) emphasize, we do not currently have such theories —though see Zylberberg, Dehaene, Roelfsema, and Sigman (2011) for some speculations.

Fodor (1975) advances CCTM+RTM as a foundation for cognitive science. He discusses mental phenomena such as decision-making, perception, and linguistic processing. In each case, he maintains, our best scientific theories postulate Turing-style computation over mental representations. In fact, he argues that our *only* viable theories have this form.

He concludes that CCTM+RTM is "the only game in town". Many cognitive scientists argue along similar lines. C.R. Gallistel and Adam King (2009), Philip Johnson-Laird (1988), Allen Newell and Herbert Simon (1976), and Zenon Pylyshyn (1984) all recommend Turing-style computation over mental symbols as the best foundation for scientific theorizing about the mind.

Computation and Representation

Philosophers and cognitive scientists use the term "representation" in diverse ways. Within philosophy, the most dominant usage ties representation to intentionality, i.e., the "aboutness" of mental states. Contemporary philosophers usually elucidate intentionality by invoking *representational content*. A representational mental state has a content that represents the world as being a certain way, so we can ask whether the world is indeed that way. Thus, representationally contentful mental states are *semantically evaluable* with respect to properties such as truth, accuracy, fulfillment, and so on. To illustrate:

- Beliefs are the sorts of things that can be true or false. My belief *that Emmanuel Macron is French* is true if Emmanuel Macron is French, false if he is not.
- Perceptual states are the sorts of things that can be accurate or inaccurate. My perceptual experience *as of a red sphere* is accurate only if a red sphere is before me.
- Desires are the sorts of things that can fulfilled or thwarted. My desire *to eat chocolate* is fulfilled if I eat chocolate, thwarted if I do not eat chocolate.

Beliefs have *truth-conditions* (conditions under which they are true),
perceptual states have *accuracy-conditions* (conditions under which they are accurate),
desires have *fulfillment-conditions* (conditions under which they are fulfilled).

In ordinary life, we frequently predict and explain behavior by invoking beliefs, desires, and other representationally contentful mental states. We identify these states through their representational properties.

Folk psychology assigns a central role to *intentional descriptions*, i.e., descriptions that identify mental states through their representational properties. Whether scientific psychology should likewise employ intentional descriptions is a contested issue within contemporary philosophy of mind.

Intentional realism is realism regarding representation. At a minimum, this position holds that representational properties are genuine aspects of mentality. Usually, it is also taken to hold that scientific psychology should freely employ intentional descriptions when appropriate....

Eliminativism is a strong form of anti-realism about intentionality. Eliminativists dismiss intentional description as vague, context-sensitive, interest-relative, explanatorily superficial, or otherwise problematic. They recommend that scientific psychology jettison representational content.

- An early example is W.V. Quine's *Word and Object* (1960), which seeks to replace intentional psychology with behaviorist stimulus-response psychology.
- Paul Churchland (1981), another prominent eliminativist, wants to replace intentional psychology with neuroscience.

Between intentional realism and eliminativism lie various intermediate positions. Daniel Dennett (1971, 1987) acknowledges that intentional discourse is predictively useful, but he questions whether mental states *really* have representational properties...
Donald Davidson (1980) espouses a neighboring *interpretivist* position. He emphasizes the central role that intentional ascription plays within ordinary interpretive practice, i.e., our practice of interpreting one another's mental states and speech acts. At the same time, he questions whether intentional psychology will find a place within mature scientific theorizing...

Computation as Formal

Classical computationalists typically assume what one might call *the formal-syntactic conception of computation* (FSC). The intuitive idea is that computation manipulates symbols in virtue of their formal syntactic properties rather than their semantic properties.

FSC stems from innovations in mathematical logic during the late 19th and early 20th centuries, especially seminal contributions by George Boole and Gottlob Frege. In his *Begriffsschrift* (1879/1967), Frege effected a thoroughgoing *formalization* of deductive reasoning. To formalize, we specify a *formal language* whose component linguistic expressions are individuated non-semantically (e.g., by their geometric shapes)...

FSC holds that *all* computation manipulates formal syntactic items, without regard to any semantic properties those items may have. Precise formulations of FSC vary... But the intuitive picture is that syntactic properties have causal/explanatory primacy over semantic properties in driving computation forward.

Fodor's article "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology" (1980) offers an early statement.
Fodor combines FSC with CCTM+RTM. He analogizes Mentalese to formal languages studied by logicians: it contains simple and complex items individuated non-semantically, just as typical formal languages contain simple and complex expressions individuated by their shapes. Mentalese symbols have a semantic interpretation, but this interpretation does not (directly) impact mental computation. A symbol's formal properties, rather than its semantic properties, determine how computation manipulates the symbol.

- In that sense, the mind is a "syntactic engine". Virtually all classical computationalists follow Fodor in endorsing FSC.

Fodor combines CCTM+RTM+FSC with intentional realism. He holds that CCTM+RTM+FSC vindicates folk psychology by helping us convert common sense intentional discourse into rigorous science. He motivates his position with a famous abductive argument for CCTM+RTM+FSC (1987: 18–20).

> !! Strikingly, mental activity tracks semantic properties in a coherent way. For example, deductive inference carries premises to conclusions that are true if the premises are true. How can we explain this crucial aspect of mental activity? Formalization shows that syntactic manipulations can track semantic properties, and computer science shows how to build physical machines that execute desired syntactic manipulations. If we treat the mind as a syntax-driven machine, then we can explain why mental activity tracks

> semantic properties in a coherent way. Moreover, our explanation does not posit causal mechanisms radically different from those posited within the physical sciences. We thereby answer the pivotal question: *How is rationality mechanically possible*?

..... a lot and I mean A LOT of yapping goes on here

Externalism about mental content

Putnam's landmark article "The Meaning of 'Meaning'" (1975: 215–271) introduced the *Twin Earth thought experiment*, which postulates a world just like our own except that H2O is replaced by a qualitatively similar substance XYZ with different chemical composition. Putnam argues that XYZ is not water and that speakers on Twin Earth use the word "water" to refer to XYZ rather than to water.

Basically

Burge concludes that mental content does not supervene upon internal neurophysiology. Mental content is individuated partly by factors outside the thinker's skin, including causal relations to the environment. This position is *externalism about mental content*.

.... Also a lot of yapping

Content-involving computation

The perceived gulf between computational description and intentional description animates many writings on CTM. A few philosophers try to bridge the gulf using computational descriptions that individuate computational states in representational terms. These descriptions are *content-involving*, to use Christopher Peacocke's (1994) terminology. On the content-involving approach, there is no rigid demarcation between computational and intentional description. In particular, certain scientifically valuable descriptions of mental activity are both computational and intentional. Call this position *content-involving computationalism*.

Footnotes

1. The label “classical” is sometimes taken to include additional doctrines beyond the core thesis that mental activity is Turing-style computation: e.g., that mental computation manipulates symbols with representational content; or that mental computation manipulates mental representations with part/whole constituency structure; or that mental computation instantiates something like the Von Neumann architecture for digital computers. Note also that the abbreviation “CCTM” is sometimes instead used as shorthand for the *connectionist computational theory of mind*.↩
2. (Johnson-Laird 2004: 187; McDermott 2001: 69; Pinker 2005: 7; Sloman 1978: 144–176).↩

Critique of Computationalism - Lecture

- Computation is not sufficient for cognition
- Syntax does not give you semantics (the so-called grounding problem)
- Grounding problem: how do arbitrary symbols get to stand for something?
  - (the problem of intentionality and aboutness)

Learning Goals

1. Understanding the impact of the Chinese Room Argument
2. Understanding the Symbol Grounding Problem
3. Understanding the different levels of Turing indistinguishability
4. Ability to critically reflect on the scope of the Chinese Room argument

Overview

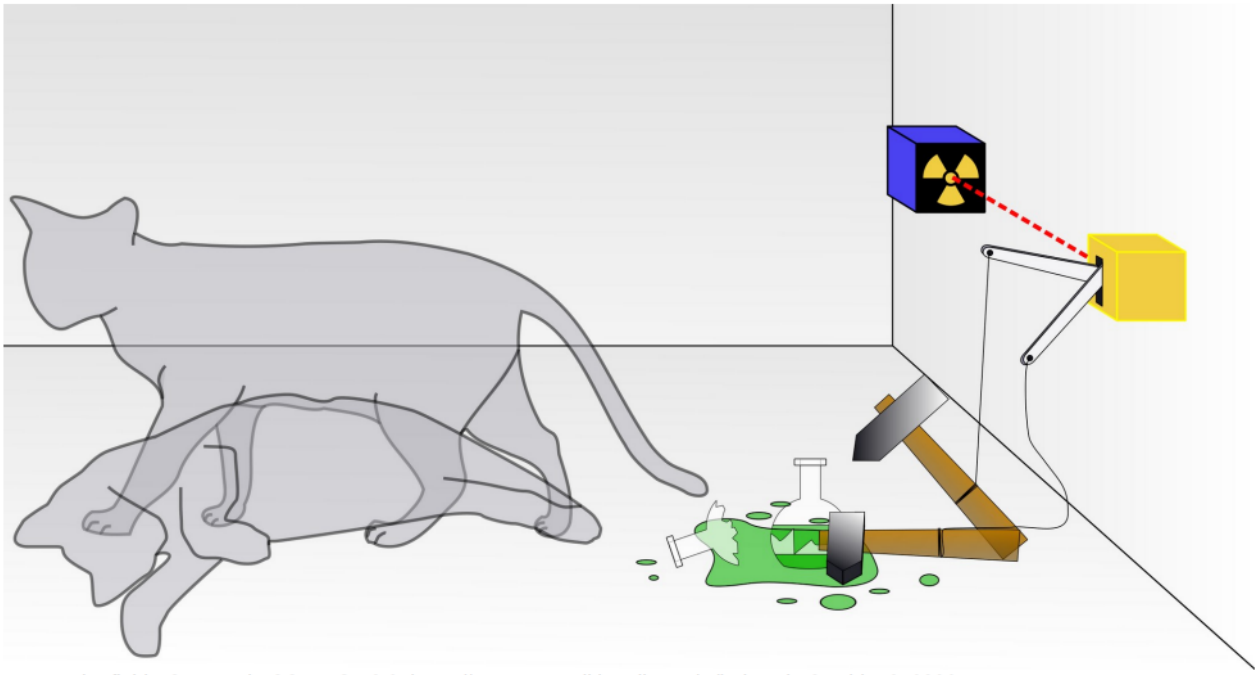
- The Chinese room thought experiment
  - What are thought experiments?
- Symbol grounding problem
  - Intentionalism – i.e. mental states are about something
- Turing indistinguishability
  - Does the Chinese room thought experiment impugn all kinds of Turing indistinguishability?

Thought experiments

- Done without leaving the armchair
- Known from both science and philosophy
- Also called *intuition pumps* (Dennett)

- Shaping the intuitions of an audience by taking something abstract and making it vivid and relatable
- Make a good one, and you will be cited forever

(Schrödinger's cat i.e.)



By Dhatfield - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=4279886

In philosophy

Hume's missing shade of blue
Conclusion?

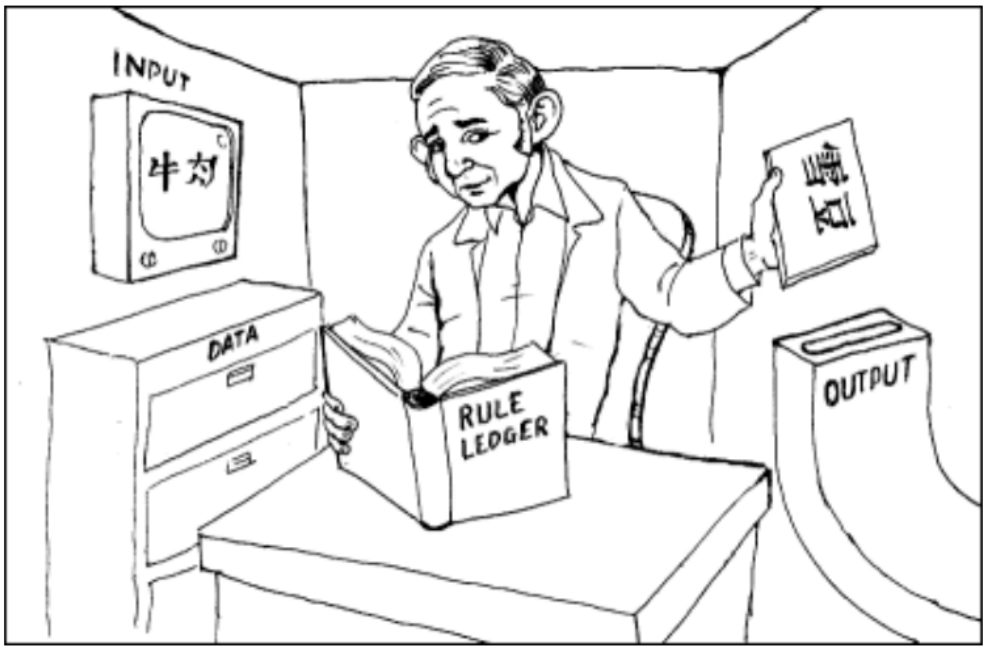
Jackson's knowledge argument
Conclusion: there are non-physical facts about colour sensation

Mary's Room: This thought experiment by Frank Jackson imagines a scientist named Mary who knows all the physical information about color but has never experienced it. The question is whether Mary learns anything new when she experiences color for the first time.

Chinese Room

We are taking something abstract ... i.e. implementation independent, systematically interpretable symbol manipulation

... and making it more vivid and relatable ...



Searle's axioms

"Axiom 1. *Computer programs are formal (syntactic).*"

"Axiom 2. *Human minds have mental contents (semantics).*"

"Axiom 3. *Syntax by itself is neither constitutive of nor sufficient for semantics.*"

Axiom 1:

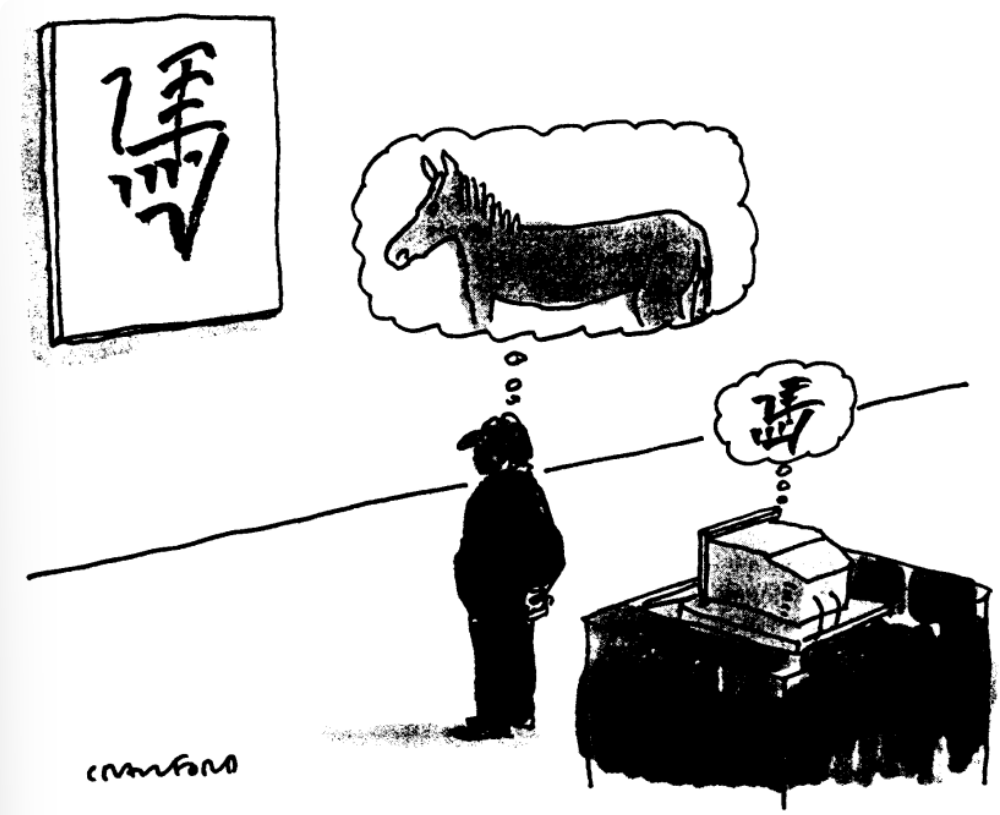
Icon: *represents* through similarity (intrinsic relation)

Index: *represents* by pointing to

A

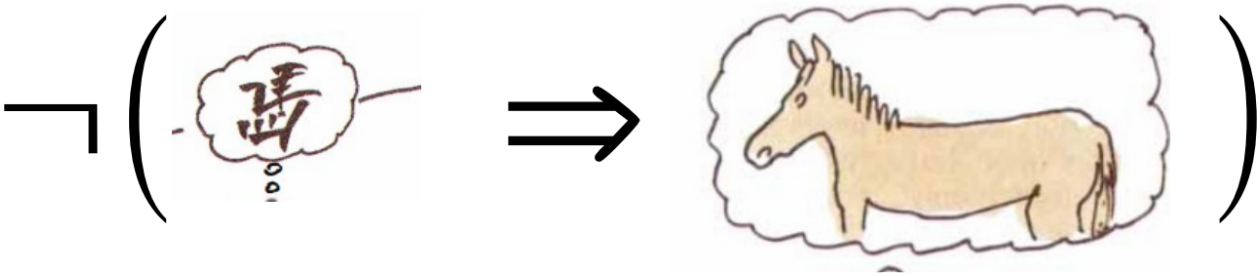
Symbol: *represents* by convention or definition (no intrinsic relation)

Axiom 2:



Computer programs are formal (syntactic).
Human minds have mental contents (semantics)

Axiom 3:



axiom (ˈæksɪəm)

n

1. (Mathematics) a generally accepted proposition or principle, sanctioned by experience; maxim
2. a universally established principle or law that is not a necessary truth: *the axioms of politics*.
3. (Logic) a self-evident statement
4. (Logic) *logic maths* a statement or formula that is stipulated to be true for the purpose of a chain of reasoning: the foundation of a formal deductive system. Compare **assumption**[4]
5. (Mathematics) *logic maths* a statement or formula that is stipulated to be true for the purpose of a chain of reasoning: the foundation of a formal deductive system. Compare **assumption**[4]

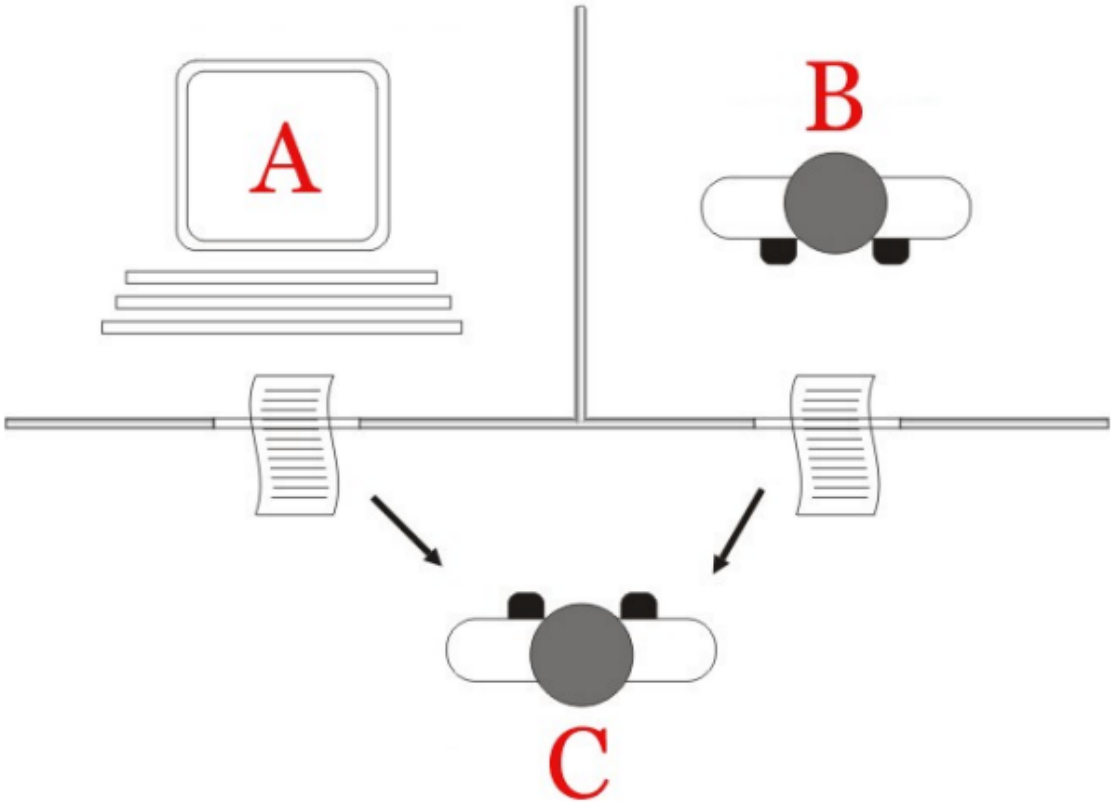
[C15: from Latin *axiōma* a principle, from Greek, from *axioun* to consider worthy, from *axios* worthy]

Searle's conclusion

"... Programs are neither constitutive of nor sufficient for minds."

Thus: C≠C

But...? Didn't "Searle in the room" "pass" the Turing Test?



In fact computer programmes had (1970s) already been passing the Turing test (within limited settings)

A restaurant script Script.png

STORY: "A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip"

QUESTION: "Did the man eat the hamburger?"

- What would an artificial intelligence equipped with scripts answer?
- What would an artificial intelligence not equipped with scripts answer?

Artificial intelligence built (1977) that could implement such scripts in answering questions like the "hamburger question" and pass the Turing test of indistinguishability (within the script settings)

Interim Summary

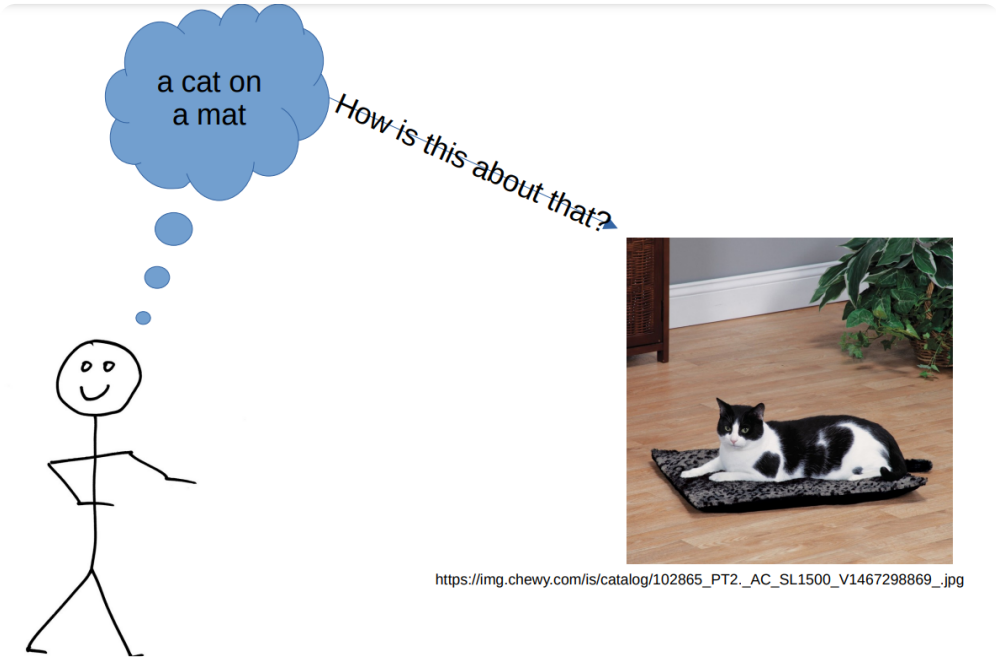
- Despite real-world success (Schank and Abelson (scripts)) and thought-experiment "success" (Searle) in obtaining Turing indistinguishability...

- Searle claims that there is no understanding and no meaning in purely formal programs, because they are not about anything, which is characteristic of (human) minds
- Next: how do the symbols get grounded?

Harnad / Symbol Grounding Problem

“The symbol grounding problem is [...] that of connecting symbols to what they are about without [my emphasis] the mediation of an external interpretation” (p. 386)

How is it that our thoughts are *about* something. How do they contain semantics?



External interpretation SYMBOLS AREN’T GROUNDED INTERNALLY

From Last Time
Now we have a programme!

What do we use the Turing test for?

- Aim of cognitive science
  - Reverse engineering of the mind
- Reverse engineering
  - Figuring out the functions and mechanics of a system we have not designed
- A good way of reverse engineering
  - Design something that is functionally indistinguishable
- Turing test
  - A test for checking whether things are functionally indistinguishable

Leibniz's law Identity of indiscernibles

$$\forall_x \forall_y [\forall F(F_x \Leftrightarrow F_y) \Rightarrow x = y]$$

In plain(er) words: for any x and y, if x and y have all the same properties, F, then x and y are identical

In other words: if it looks like a duck, sounds like a duck, smells like a duck, quacks like a duck, then it is a duck

Turing indistinguishability

T4 indistinguishability

"T4: Symbolic, sensorimotor and neuromolecular *indistinguishability*"

That is, a perfectly reverse engineered duck

Knowing everything about this T4-duck, would mean you would know everything about ducks - Leibniz's law Leibniz's law Identity of indiscernibles

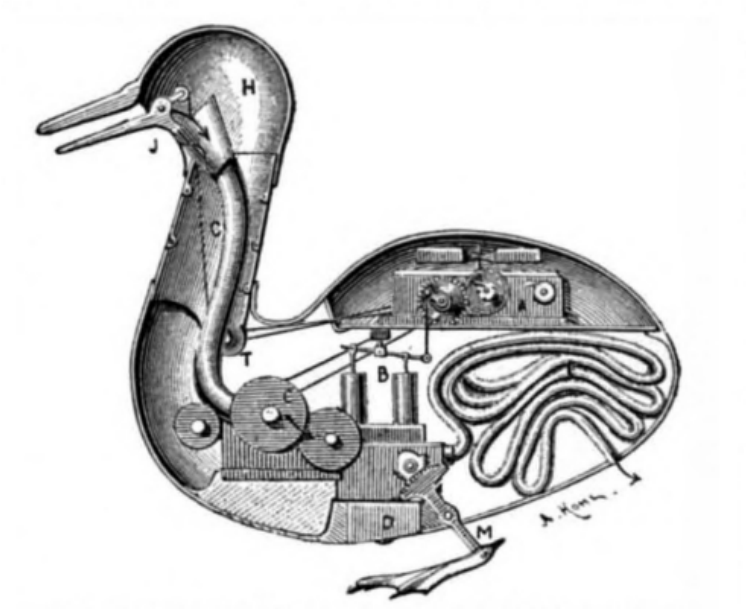
T3 indistinguishability

T3: Symbolic and sensorimotor indistinguishability

That is, a functionally reverse engineered duck

Would likely teach you something about real ducks

This introduces structural demands, but not structural specificity



INTERIOR OF VAUCANSON'S AUTOMATIC DUCK.

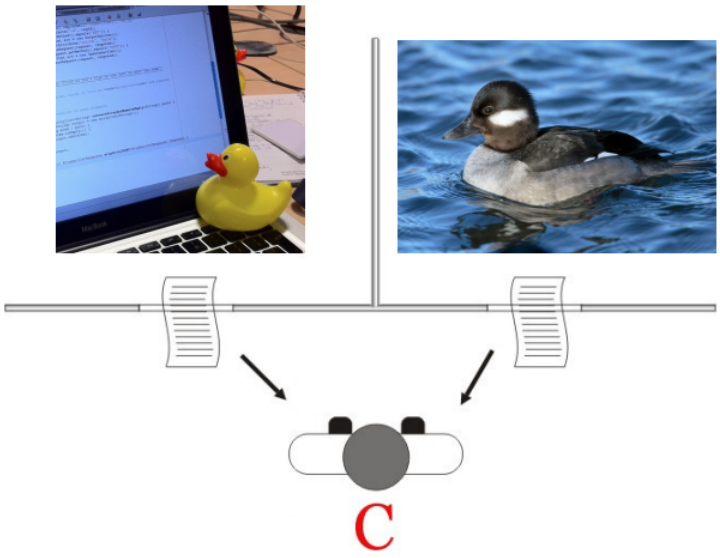
A, clockwork; *B*, pump; *C*, mill for grinding grain; *F*, intestinal tube; *J*, bill; *H*, head; *M*, feet.

T2 indistinguishability

T2: Symbolic indistinguishability

That is, a limited range of functionality is indistinguishable

This is not tied to structure, and is therefore implementation independent



Let's get back to Searle and some of the replies he considers

The Replies

c) The systems reply
g) The brain simulator reply
f) The robot reply

The systems reply

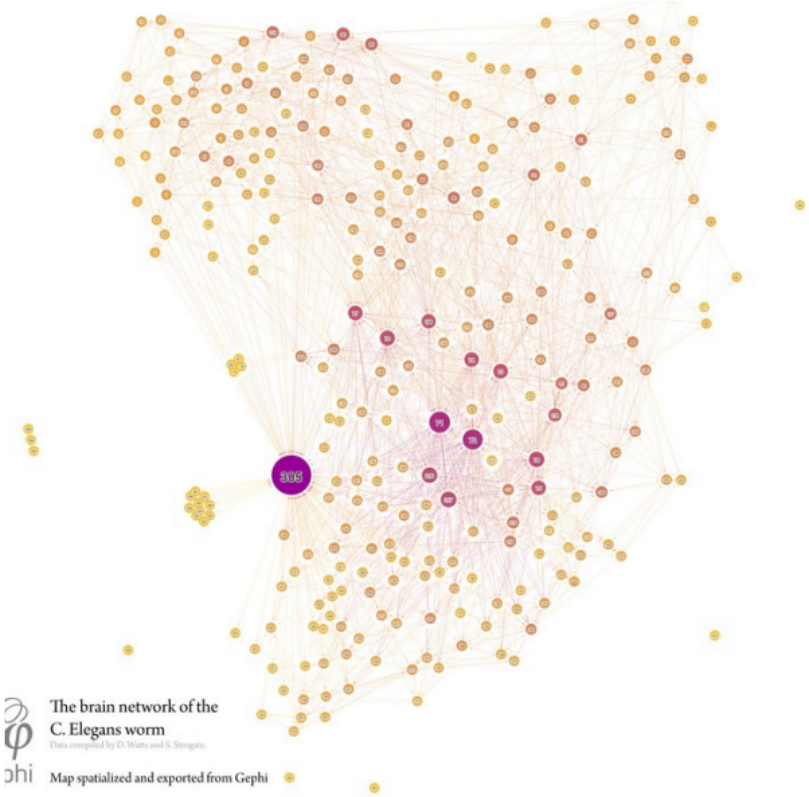
It might be the case that Searle does not understand Chinese, but the room as a whole does

- Compare with a single (language) neuron relative to the whole brain

Searle's reply to the systems reply
Saying that the room understands Chinese is analogous to claiming that he (Searle) understands Chinese, if he memorizes all the symbols and all the rules

The brain simulator reply

If the operations of the brain of a Chinese speaker were simulated, such a system would understand Chinese, just like any Chinese speaker does



Searle's reply to the brain simulator reply

"Computer simulations of brain processes provide models of the formal aspects of these processes. But [...] [t]he computational model is no more real than the computational model of any other natural phenomenon" p. 29

- see also On why people think models of the mind, actually think

The robot reply

If the computer programme was causally hooked up with the rest of the world, then it would acquire semantics (it would be able to ground symbols by causally interacting with the world)

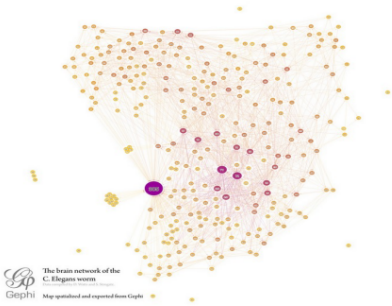
Searle's reply to the robot reply

Imagining the Chinese room inside the robot does not add semantics, even if the robot interacts with the world

- (*We'll consider this in more detail in the predictive processing lecture*)

Turing indistinguishability again

T2 indistinguishability



T3/T4

T3/T4

Note: not implementation-independent

The upshot of the symbol grounding problem

- Insisting on *implementation independence*, means that you can only design T2-indistinguishable systems
  - which will not have intrinsic semantics and thus not reflect cognition, which seems to be *about* something
- To get semantics, you need at least T3-indistinguishability
  - losing implementation independence
- The Chinese room argument does not work against Turing indistinguishability as such
  - only against limited functional mimicry (T2)

Two potential solutions to the symbol grounding problem

- Connectionism
  - Making a system that is functionally identical to (human) minds (brain simulation)
- Predictive processing
  - Understanding the human mind as a system whose function and structure enables it to make causal inferences about the world

Summary

- The Chinese Room Argument shows that implementation independence and semantics are at odds with one another

- This is because human cognition seems to be *about* something, whereas computation is purely formal
- The symbol grounding problem is how formal signs get their meaning. (How is the sign grounded?)
- Accepting some implementation *dependence* (T3/T4) may allow for building systems that can ground symbols

Next Lecture

Week 15 – Connectionism

- Computation happens below the representational level in a connected network of units (e.g. neurons)
- Semantics emerges from the network
- The algorithm level is not independent from the implementation level

Harnad - Computation is just interpretable symbol manipulation; cognition isn't

Abstract

> "... But even after computers and computation have been successfully distinguished from other kinds of things, mental states will not just be the implementations of the right symbol systems, because of the symbol grounding problem: The interpretation of a symbol system is not intrinsic to the system; it is projected onto it by the interpreter. This is not true of our thoughts. We must accordingly be more than just computers. My guess is that the meanings of our symbols are grounded in the substrate of our robotic capacity to interact with that real world of objects, events and states of affairs that our symbols are systematically interpretable as being about."

... we must first agree about what *computation* is (*cognition* will be a much harder one to agree on, but being all Cartesian cognizers ourselves, let us settle for an ostensive definition on that for now).

Let me declare right away that I subscribe to what has come to be called the Church/Turing Thesis (CTT) (Church 1956), which is based on the converging evidence that all independent attempts to formalize what mathematicians mean by a "computation" or an "effective procedure," even when they have looked different on the surface, have turned out to be equivalent (Galton 1990).

Symbol Manipulation Systems and the Church/Turing Thesis

According to all these notational variants on what we might as well call *Turing Machines*, computation is just the manipulation of symbol tokens on the basis of their shapes (Turing 1990).

... everything that mathematicians, logicians and computer scientists have done so far by means of logical inference, calculation and proof can be done by the machine. I say "so far," because it is still an open question whether people can "compute" things that are not computable in this formal sense: If they could, then CTT would be false.

The Thesis is hence not a Theorem, amenable to proof, but an *inductive conjecture*[1] supported by evidence; yet the evidence is about formal properties, rather than about physical, empirical ones.

There is a natural generalization of CTT to *physical systems* (CTTP). According to the CTTP, everything that a discrete physical system can do (or everything that a continuous physical system can do, to as close an approximation as we like) can be done by computation.

The CTTP comes in two dosages:

- A Weak and a Strong CTTP, depending on whether the thesis is that all physical systems are formally equivalent to computers *or* that they are just computers.

Nothing rides on this until we come to the property of implementation-independence, below, and then the distinction will turn out to raise substantive problems for Computationalism (the thesis that cognition is just a form of computation, C=C), a thesis that will also come in a Weak and Strong form.

It is accordingly important to point out that I (Harnad, red) do accept the Church-Turing Thesis, in both its formal and physical versions (CTT and CTTP), yet I too will be arguing against C=C.

Systematic Interpretability

There are still two important components of the definition of computation that I have left out.

The first of these is controversial and is not usually regarded as part of the definition:

- Formal computation is clearly symbol manipulation, with the operations on the symbols (read, write, move, halt) being based, as already stated, on the shapes of the symbols.

Such shape-based operations are usually called "*syntactic*" to contrast them with "*semantic*" operations, which would be based on the meanings of symbols, rather than just their shapes

- see Searle - Is the Brain's Mind a Computer Program for more on syntactic vs semantic

Meaning does not enter into the definition of formal computation.

So although it is usually left unstated, it is still a criterial, if not a definitional property of computation that the symbol manipulations must be semantically interpretable -- and not just locally, but globally:

- All the interpretations of the symbols and manipulations must square systematically with one another, as they do in arithmetic, at the level of the individual symbols, the formulas, and the strings of formulas.
- It must all make systematic sense, in whole and in part (Fodor & Pylyshyn 1988).

This criterion of semantic interpretability (which has been dubbed the "*cryptographers constraint*," because it requires that the symbol system should be decodable in a way that makes systematic sense) is not a trivial one to meet:

- It is easy to pick a bunch of arbitrary symbols and to formulate arbitrary yet systematic syntactic rules for manipulating them, but this does not guarantee that there will be any way to interpret it all so as to make sense (Harnad 1994b).

Computation: Trivial Vs. Nontrivial

Trivial vs. Nontrivial Symbol Systems:

- *Trivial* symbol systems are those where the symbols can be arbitrarily interpreted without affecting the overall coherence or meaning of the system. For example, a system where "0" represents "Life is like a bagel" and "1" represents "Life is not like a bagel" is trivial because swapping the interpretations doesn't affect the overall coherence.
- *Nontrivial* symbol systems are those where the interpretation of symbols is *not arbitrary* and swapping interpretations may result in a loss of coherence or meaning. For example, in English, swapping the interpretations of "true" and "false" or "red" and "green" would disrupt the coherence of the language.

Computability and Interpretation:

- Harnad suggests that generating uninterpretable symbol systems is a form of trivial computation because there is no meaningful interpretation that can be derived from them. Nontrivial computation, on the other hand, involves symbol systems that can be systematically interpreted.
- He emphasizes that meaningful interpretation is a crucial aspect of nontrivial symbol systems and distinguishes them from trivial ones.

Physical Implementation of Symbol Systems:

- Harnad discusses how a computer serves as the physical implementation of a symbol system. In a computer, states and state-sequences represent the interpretable objects.
- He distinguishes between abstract idealizations of symbol systems, such as the Universal Turing Machine, and concrete physical realizations like digital computers.

Implications for Classification of Systems:

- Harnad suggests that if the distinction between nontrivial and trivial symbol systems can be formally worked out, certain entities like walls might be excluded from the class of computers because they only implement trivial computations.

Implementation Independence

So we are interested only in nontrivial computation. That means symbols, manipulated on the basis of their shapes only, but nevertheless amenable to a systematic interpretation. Symbol systems that are meaningful, in other words.

- Summarizes the foregoing section

The shapes of the symbol tokens must be arbitrary. Arbitrary in relation to what? In relation to what the symbols can be interpreted to mean. For example,

- in formal Peano arithmetic, the equality symbol "=" is manipulated purely on the basis of its shape, and its shape bears no physical relation to the property of "equality"
  - it neither resembles it nor is it causally connected to it.
  - The same is true of "3," which neither resembles "threeness" nor is causally connected to it in the world.

This is a natural place to point out that the symbols of natural language likewise have this property of arbitrariness in relation to what they mean (what Saussure[2] called "l'arbitraire du signe).

... We need only note that, among symbolic systems, those that are doing nontrivial computation are rare indeed. We may need a successful human interpretation to prove that a given system is indeed doing nontrivial computation, but that is just an epistemic matter. If, in the eye of God, a potential systematic interpretation exists, then the system is computing, whether or not any Man ever finds that interpretation.

A cat on a mat can be interpreted as meaning a cat on the mat, with the cat being the symbol for cat, the mat for mat, and the spatial juxtaposition of them the symbol for being on.

- Why is this not computation?
  - Because the shapes of the symbols are not arbitrary in relation to what they are interpretable as meaning, indeed they are precisely what they are interpretable as meaning

Again, this may seem a trivial case to single out, but it grades into more problematic cases: those in which the "symbols" physically resemble or are causally connected to what they mean. The simple operations of a Turing Machine are not based on any direct connection between the symbols and their meanings

- They are purely formal or *syntactic* and therefore universal

Another way of characterizing the arbitrariness of the shapes of the symbols in a formal symbol system is as "**implementation independent**":

Completely different symbol-shapes could be substituted for the ones used, yet if the system was indeed performing a computation, it would continue to be performing the same computation if the new shapes were manipulated on the basis of the same syntactic rules.

> The power of computation comes from the fact that neither the notational system for the symbols nor the particulars of the physical composition of the machine are relevant to the computation being performed.

Turing Indistinguishability

So let us write out long-hand what computation has turned out to be: **implementation-independent, systematically interpretable, symbol manipulation.**

Now if the interpretation is mentalistic and not merely semantic, it becomes even more irresistible, forever confirming itself (by definition, in virtue of its systematic semantic interpretability); I have called this the "hermeneutic hall of mirrors" (Harnad 1990 b, c, Hayes et al.1992).

According to Turing, if there was a person whom we could not tell apart from other people in any way (for years and years, say), then we would have no nonarbitrary basis for concluding that he had no mind if we were informed that he was a machine.
- This could be construed as encouragement to succumb to the hermeneutic power of semantic interpretability -- particularly because, to rule out biases based on appearance, the Turing Test was formulated as involving a pen-pal, who communicated with symbols only.
- But one can give Turing credit for more sense than that, and interpret him as arguing that only a fool would presume to try to distinguish between functional indistinguishables (Harnad 1992 b, 1994 b).

The Symbol Grounding Problem

Basically...

Relates to how symbols get their meaning and therefore relates what meaning is

Meaning cannot be inferred

The Chinese room understands symbols (syntax) but doesn't understand what the term is *about* (semantics) = no external validity

Computation is *syntactic information processing*

Real people do *semantic information processing*

So I see Turing as championing machines in general that have functional capacities indistinguishable from our own, rather than computers and computation in particular.

- Yet there are those who do construe Turing's Test as support for C=C.
- They argue: Cognition is computation. Implement the right symbol system -- the one that can pass the penpal test (for a lifetime) -- and you will have implemented a mind.

Unfortunately, the proponents of this position must contend with Searle's (1980) celebrated Chinese Room argument, in which he pointed out that any person could take the place of the penpal computer, implementing exactly the same symbol system, without understanding a word of the penpal correspondence.

- Since computation is implementation-independent, this is evidence against any understanding on the part of the computer when it is implementing that same symbol system.

But, as I suggested, Searle's Argument does not really impugn Turing Testing (Harnad 1989); it merely impugns the purely symbolic, pen-pal version of the Turing Test, which I have called T2. It leaves the robotic version (T3) -- which requires Turing-indistinguishable symbolic and sensorimotor capacity -- untouched (just as it fails to touch T4: symbolic, sensorimotor and neuromolecular indistinguishability).

- See Turing indistinguishability

So only a pure symbol system is vulnerable to Searle's Argument, and it is not hard to see why. I have dubbed the reason the "**Symbol Grounding Problem**" (Harnad 1990 a,1993 c): No one knows what cognition is, but we know that cognizers do it.

Long one now:
A string of symbols such as "2 + 2 = 4" or "the cat is on the mat," generated by a symbol system, is an instance of nontrivial computation if it is systematically interpretable as meaning what "2 + 2 = 4" and "the cat is on the mat" mean.

- But that meaning, as stated earlier, is not contained in the symbol system.

- The system is merely syntactic, manipulating meaningless symbols on the basis of shape-based rules, and the shapes of the symbols are arbitrary in relation to what they are interpretable as meaning.'

Looking for meaning in such a system is analogous to looking for meaning in a Chinese/Chinese dictionary when one does not know any Chinese:

- All the words are there, fully defined; it is all systematic and coherent. Yet if one looks up an entry, all one finds is a string of meaningless symbols by way of definition,
- if one looks up each of the definienda in turn, one just finds more of the same.

The search is ungrounded. It is all systematically interpretable to someone who already knows some Chinese, but in and of itself it is meaningless and leads only to an infinite regress.

The critical divergence point between computation and cognition: I have no idea what my thoughts are, but there is one thing I can say for sure about them:

- They are thoughts about something, they are meaningful, and they are not about what they are about merely because they are systematically interpretable by you as being about what they are about.
- They are about them autonomously and directly, without any mediation

The symbol grounding problem is accordingly that of connecting symbols to what they are about without the mediation of an external interpretation (Harnad 1992 d, 1993 a).

- The "external interpretation" is the key here.
  - Kinda ish maybe: How do we define meaning without using meaning in the first place

One solution that suggests itself is that T2 needs to be grounded in T3:

- Symbolic capacities have to be grounded in robotic capacities.

Many skeptical things could be said about a robot who is T3-indistinguishable from a person (including that it may lack a mind), but it cannot be said that its internal symbols are about the objects, events, and states of affairs that they are about only because they are so interpretable by me, because the robot itself can and does interact, autonomously and directly, with those very objects, events and states of affairs in a way that coheres with the interpretation.

- It tokens "cat" in the presence of a cat, just as we do, and "mat" in the presence of a mat, etc. And all this at a scale that is completely indistinguishable from the way we do it, not just with cats and mats, but with everything, present and absent, concrete and abstract.

That is guaranteed by T3, just as T2 guarantees that your symbolic correspondence with your T2 pen-pal will be systematically coherent.

But there is a price to be paid for grounding a symbol system:

- It is no longer just computational! At the very least, sensorimotor transduction is essential for robotic grounding, and transduction is not computation.

Cognition: Virtual Vs. Real

Or is it? Let us recall the two versions of Church's Thesis introduced earlier: There was the purely formal one, CTT, and the physical one, CTTP, and the latter came in a Weak and Strong form.

There is no problem with the Weak CTTP, because it merely claims that every physical system is formally equivalent to a Turing Machine, and that would be true of a transducer as well, just as it would be of an airplane, a furnace, or a solar system:

- All of these can be simulated, state-for-state, by a computer, to as close an approximation as one likes.

... All of this should be quite obvious. A bit less obvious is the equally valid fact that a virtual pen-pal does not think (or understand, or have a mind) -- because he is just a symbol system systematically interpretable as if it were thinking (understanding, mentating).

The truth of the Weak CTTP would guarantee that such simulations are possible, and it would strongly imply that computational modeling was an excellent way of arriving at an understanding of physical systems (Searle (1980) called this "Weak AI" (see Can a Machine Think)), but it would not support C=C ("Strong AI") Why?

- Because even if the simulated T3 robot captured (i.e., simulated) every relevant property of cognition, it would still be just an ungrounded symbol system.
  ...

Differential Equations and Implementation Dependence

What about the Strong CTTP, according to which a plane is a computer? Well, either this is wrong, or it is uninteresting (lol)

But I actually think the Strong CTTP is wrong, rather than just vacuous, because it fails to take into account the all-important **implementation-independence** that does distinguish computation as a natural kind: *For flying and heating, unlike computation, are clearly not implementation-independent.*

- The pertinent invariant shared by all things that fly is that they obey the same sets of differential equations, not that they implement the same symbol systems (Harnad 1993 a)

So much for strong CTTP. What about C=C? A weak version, according to which cognition can be simulated -- to as close an approximation as one likes, by computation -- is really just a variant of the Weak CTTP:

- Why would a nondualist expect that cognition would differ in this respect from any other physical process (flight, heat, gravity), all likewise simulable by computation?

But Strong Computationalism, according to which every implementation of the right symbol system would cognize, is either wrong in exactly the same way the Strong CTTP is wrong (or vacuous)...

Reading Questions

1. In your own words, what does Harnad mean when he says that the Church/Turing Thesis is an "inductive conjecture"?
2. What does it mean when Harnad states that he accepts CTT and CTTP, but not C=C?
3. What is the criterion of semantic interpretability? And why is it important?
4. Based on your own understanding and/or quotations from the text, what is your best definition of the “symbol grounding problem”

Footnotes

1. By calling the Church-Turing Thesis an "inductive conjecture," Harnad means that it's not a theorem that can be proven definitively, but rather a hypothesis based on observed evidence. (Mr. GPT). Basically the induction problem; there might be a problem which is not computable but it is yet unknown↩
2. Tog Dog↩

Searle - Is the Brain's Mind a Computer Program

> // No a program merely manipulates symbols, whereas the brain attaches meaning to them

Can a Machine think?

Humans are a special biological kind of machine, and humans can think, so of course machines can think. Maybe it will turn out to be impossible to make one, but we don't know yet.

In recent decades however the question has been posed differently.

- Could a machine think by virtue of implementing a computer program?
- Is the program itself constitutive of thinking?

These are very different questions because they do not revolve around the physical, causal attributes of actual or physical systems but rather about the abstract, computational properties of formal, computer programs that can be implemented in any sort of substance at all

... So the goal is to design programs that will simulate human cognition in such a way as to pass the Turing test. What is more, such a program would not merely be a model of the mind; it would literally be a mind, in the same sense that a human mind is a mind.

Searle distinguishes between Strong AI and Weak AI

- Strong AI claims that thinking is merely formal symbol manipulation, i.e. what a computer does
  - "The mind is to the brain as the program is to the hardware"

Strong AI is unusual among theories of the mind in at least two respects

1. It can be stated clearly
2. It admits of a simple and decisive refutation

Refutation (The Chinese Room argument)
Consider a language you don't understand, like Chinese. The symbols of Chinese writing look like meaningless squiggles. Suppose you are placed in a room containing baskets full of Chinese symbols. Suppose also that a English rule book is given for matching Chinese symbols with Chinese symbols. The rules identify the symbols entirely by their shapes.

Imagine people outside the room who understand Chinese hand in small bunches of symbols and that in response I manipulate the symbols according to the rule book. The rule book is the "computer program", the people who wrote it are the "programmers" and I am the "computer". The baskets full of symbols are the "data base", the small bunches of symbols that are handed in to me are "questions" and the bunches handed out are "answers" I manipulate symbols but attach no meaning.png

If I do not understand Chinese solely on the basis of running a computer program for understanding Chinese, then neither does any other digital computer solely on that basis. Digital computers merely manipulate formal symbols according to rules in the program.

This simple argument is decisive against the claims of Strong AI

Axiom 1: *Computer programs are formal (syntactic)*

- They can be run on an indefinite variety of hardwares
- A digital computer processes information by first encoding it in the symbolism that the computer uses (1s ad 0s) and then manipulating the symbols through a set of precisely states rules
    - These rules constitute the program

The astonishing thing about computers is that any information that can be stated in a language can be encoded in such a system, and any information processing task that can be solved by explicit rules can be programmed

Two Further Points

1. symbols and programs are purely abstract notions
2. symbols are manipulated without references to any meanings
    - The symbols can stand for anything the creator/user wants

- The program has syntax but no semantics

Axiom 2: *Human minds have mental contents (semantics)*

Axiom 3: *Syntax by itself is neither constitutive of nor sufficient for semantics*

From these premises (axioms) it follows that

Conclusion 1: *Programs are neither constitutive of nor sufficient for minds*

- (another way of saying that strong AI is false)

From the fact that a system can be simulated by formal symbol manipulation and the fact that it is thinking (the brain) it does not follow that thinking is equivalent to formal symbol manipulation

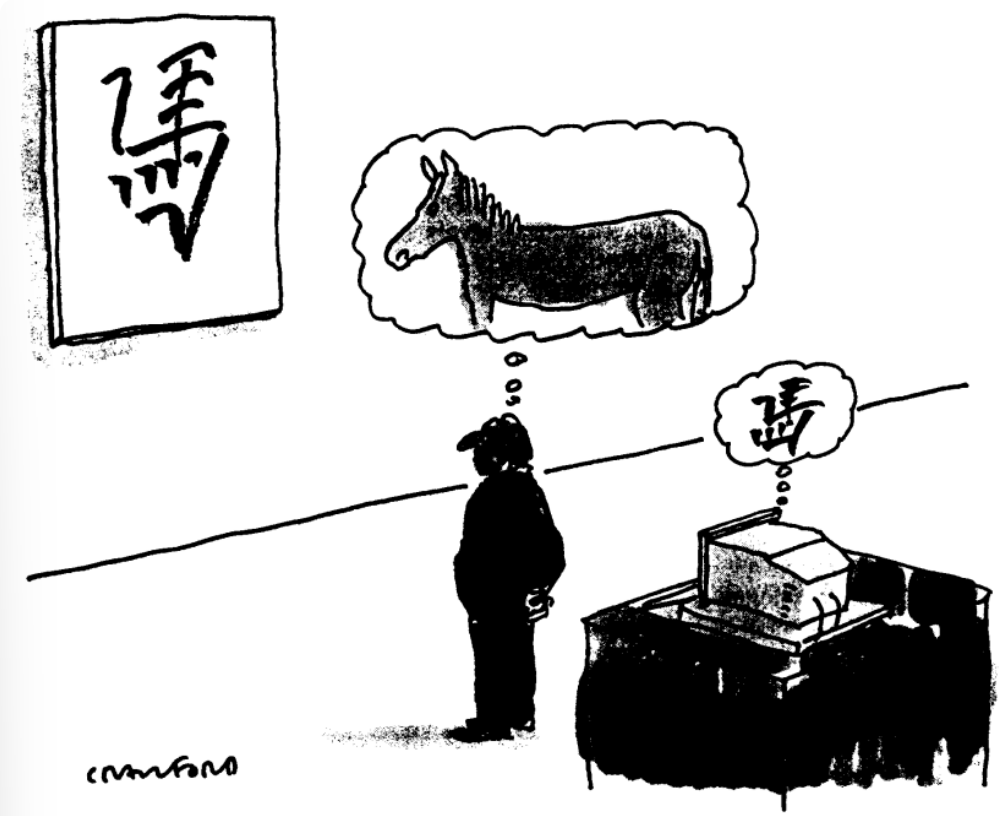
Searle is not excluding the possibility that other systems than our biological ones can think, he's just saying we not have yet discovered/invented such

Further Still

Although the results are still modest, these "parallel distributed processing" or "connectionist" models raise useful questions about how complex, parallel network systems like those in brains might actually function in the production of intelligent behavior.

Any function that can be computed on a parallel machine can be computed on a serial machine. Parallel processing then, does not afford a way around the Chinese Room argument.

> !! You can't get semantically loaded though contents from formal computations alone, whether they are done in serial or in parallel; that is why the Chinese Room argument refutes strong AI in any form



Computer programs are formal (syntactic).
Human minds have mental contents (semantics)

Many People Who

On the topic of what causes consciousness in the brain:

"Brains are specific biological organs, and their specific biochemical properties enable them to cause consciousness and other sorts of mental phenomena. Computer simulations of brain processes provide models of the formal aspects of these processes. But the simulation should not be confused with duplication. The computational model of mental processes is no more real than the computational model of any other natural phenomenon"

All mental phenomena, then, are caused by neurophysiological processes in the brain. Hence,

Axiom 4: *Brains cause minds*

Conclusion 2: *Any other system capable of causing minds would have to have causal powers (at least) equivalent to those of brains*

- this conclusion says nothing about the mechanisms

Conclusion 3: *Any artifact that produced mental phenomena, any artificial brain, would have to be able to duplicate the specific causal powers of brains, and it could not do that just by running a formal program*

Conclusion 4: *The way that human brains actually produce mental phenomena cannot be solely by virtue of running a computer program*

Common Objections to the Chinese Room

1. In the Chinese room you understand Chinese, even though you don't know it. It is, after all, possible to understand something without knowing that one understands it.
2. You don't understand Chinese, but there is an (unconscious) subsystem in you that does. It is, after all, possible to have unconscious mental states, and there is no reason why your understanding of Chinese should not be wholly unconscious.
3. You don't understand Chinese, but the whole room does. You are like a single neuron in the brain, and just as such a single neuron by itself can- not understand but only contributes to the understanding of the whole system, you don't understand, but the whole system does.
4. Semantics doesn't exist anyway; there is only syntax. It is a kind of prescientific illusion to suppose that there exist in the brain some mysterious "mental contents," "thought processes" or "semantics." All that exists in the brain is the same sort t of syntactic symbol manipulation that goes on in computers. Nothing more
5. You are not really running the computer program-you only think you are. Once you have a conscious agent going through the steps of the program, it ceases to be a case of implementing a program at all.
6. Computers would have semantics and not just syntax if their inputs and outputs were put in appropriate causal relation to the rest of the world. Imagine that we put the computer into a robot, attached television cameras to the robot's head, installed transducers connecting the television messages to the computer and had the computer output operate the robot's arms and legs. Then the whole system would have a semantics.
7. If the program simulated the operation of the brain of a Chinese speaker, then it would understand Chinese. Suppose that we simulated the brain of a Chinese person at the level of neurons. Then surely such a system would understand Chinese as well as any Chinese person's brain.

All these arguments are inadequate
The Room argument rests on the distinction between the formal symbol manipulation that is done by the computer and the mental contents biologically produced in the brain (syntax / semantics)

Number 3 (c) is the most widely help objection which Searle addresses (the systems reply)

The Systems Reply

> "Arguments from analogy are notoriously weak, because before one can make the argument work, one has to establish that the two cases are truly analogous"

- s/o Paley's Watch

...

The Churchlands' analogy between syntax and electromagnetism, then, is confronted with a dilemma; either the syntax is construed purely formally in terms of its abstract mathematical properties, or it is not.

- If it is, then the analogy breaks down, because syntax so construed has no physical powers and hence no physical, causal powers.
- If, on the other hand, one is supposed to think in terms of the physics of the implementing medium, then there is indeed an analogy, but it is not one that is relevant to strong AI.

Because the Points

On why people think a computer model of thought processes would actually think:

Part of the answer is that people have inherited a residue of behaviorist psychological theories of the past generation.

- The Turing test enshrines the temptation to think that if something behaves as if it had certain mental processes, then it must actually have those mental processes.

And this is part of the behaviorists' mistaken assumption that in order to be scientific, psychology must confine its study to externally observable behavior.

- Behaviorism

Paradoxically, this residual behaviorism is tied to a residual dualism. Nobody thinks that a computer simulation of digestion would actually digest anything, but where cognition is concerned, people are willing to believe in such a miracle because they fail to recognize that the mind is just as much a biological phenomenon as digestion.

- The mind, they suppose, is something formal and abstract, not a part of the wet and slimy stuff in our heads.

The polemical literature in AI usually contains attacks on something the authors call dualism, but what they fail to see is that they themselves display dualism in a strong form, for unless one accepts the idea that the mind is completely independent of the brain or of any other physically specific system, one could not possibly create minds just by designing programs.

> "It is best to see strong AI as one of the last gasps of this antiscientific tradition, for it denies that there is anything essentially physical and biological about the human mind"

- the mind, according to strong AI, is independent of the brain, and capable of running on any hardware

The important point is that simulation is not the same as duplication, and that fact holds as much import for thinking about arithmetic as it does for feeling angst. The point is not that the computer gets only to the 40-yard line and not all the way to the goal line.

The computer doesn't even get started. It is not playing that game.

Reading Questions

1. What definition of a "mind" do the proponents of Strong AI (referred to as "[a] fair number of researchers in AI") hold according to Searle?
2. Clearly, Searle does not agree with the Strong AI proponents, but what is then his definition of "minds"?
3. The different strands of computationalism make different claims about syntax, semantics, and their mutual relationship. The "formal-syntactic conception of

computation" (FSC; see Rescorla, 2015) view, for example, presupposes that we can have syntax without semantics. But can we have semantics without syntax? Do any of the computationalist strands you have learned about so far allow that?

4. Does Searle claim that computers cannot think? How so?
5. Do parallel distributed and/or connectionist models circumvent the Chinese Room argument (according to Searle)?
6. Which one of the seven objections against the Chinese Room argument (a-g) do you find the most convincing? Why?
7. Searle refutes the belief that computers would think as "residue of behaviourist psychological theories." But would Skinner have agreed that "if something behaves as if it had certain mental processes, then it must actually have these"

Everything, everywhere, all at once

Philosophy of Science

Philosophical foundations

- What can we know?
  - Skepticism
- How do we know?
  - Sensation
  - Reason
- What kinds of knowing?
  - analytic, synthetic, *a priori*, *a posteriori* Philosophical statements
  - manifest image vs scientific image

Learning goals:

- Understand the historical background for the modern philosophy of science (1900s onwards)
  - e.g.: dualism, empiricism and idealism
- Understand what is/was regarded as justified beliefs
  - analytic, synthetic, a priori, a posteriori
- Appreciate the relevance to today's philosophy and today's science

Can we acquire knowledge through the senses

Descartes

- Methodological doubt
  - senses cannot be trusted
  - **certain knowledge** can only be obtained about (one's own) mind(s)

Hume

- The senses are **the only way**, we can acquire knowledge about the world
  - however, this limits quite considerably what we can know about the world
    - e.g. we cannot know of causes and effects

Kant

- Yes, but we need to understand how our intuitions, i.e. our minds, **structure** our sensations
  - Relate Kant's Copernican revolution to Kuhn and also to predictive processing

Searle

- Yes, but we need to allow for **imperceptible** entities (scientific image) that explain how **perceptible** entities manifest
  - Relate to the T's we see in positivism and falsificationism

Science, non science, pseudoscience

- The demarcation problem: how do we differentiate science from non-science and especially how do we differentiate it from pseudoscience?
  - E.g.: why is intelligent design not a serious rival of evolutionary theory?
  - E.g.: why is research into precognition seen as dubious?

Learning goals:

- Acquiring an understanding of how science and pseudoscience can be differentiated, using creationism and evolutionary theory as examples
- Understanding some of the vulnerabilities of our own field – the importance of good practices, using precognition research as an example

Loose definitions

- Scientific statements
  - About the natural world
  - Used to explain how the world works
  - *Are at least in principle testable*
  - Science can become pseudoscience (e.g. phrenology)
- Pseudoscientific statements
  - Tries to look like science
  - The explanation given supports some other agenda
  - *Are not testable, or do not recognize refuting statements*
  - Can pseudoscience become science?

The barrenness of theory without assumptions

T1: living things are the product of intelligent design
T2: living things are the product of random physical processes
T3: living things are the product of evolutionary selection

What O's do these T's imply, if any

Duhem's thesis

Duhem's thesis also called Duhem-Quine

> Typically, T does not deductively imply O; rather, it is T&A that deductively implies O (here, T is a theory, O is an observation statement, and A is a set of of auxiliary assumptions) - Sober, p. 49

importantly, the auxiliary assumptions should be independently testable

Summary

Thus to do science

- we need general theories, T, that together with auxiliary assumptions, A, allow us to formulate expected observations, O. We then conduct experiments or observations that should reveal whether O is the case

Demarcation of science: logical positivism

- Creating a sharp boundary between the *context of discovery* and the *context of justification*
- The only meaningful statements are observation statements or theoretical statements that can be translated to observation statements (Verifiability criterion)
- All other statements are meaningless

Learning goals:

- Understanding how logical positivism separates meaningful from meaningless sentences
- Understanding the implications for philosophy, science and especially, psychology (a component of cognitive science)
- Understanding the problems facing logical positivism, e.g. verification of general sentences and the Quine-Duhem thesis

Criterion of meaning

a T-statement is only meaningful if (an) O statement(s) follow(s) from it

or from a Epistemological View:

- how is T to be verified?

Thus if we cannot specify, i.e., deduce O from T:
$T \rightarrow O$
then T is meaningless

Stay on the lookout for T's that cannot be operationalized!

HOW TO INVESTIGATE MENTAL CONCEPTS?

- Carnap - Psychology in Physical Language

$P_1 \Leftrightarrow P_2$
this is a **translation** (providing the ingredients for an operationalization) from psychological to physical language

P2 a dispositional statement (logical **behaviourism**)

Demarcation of science: Falsificationism

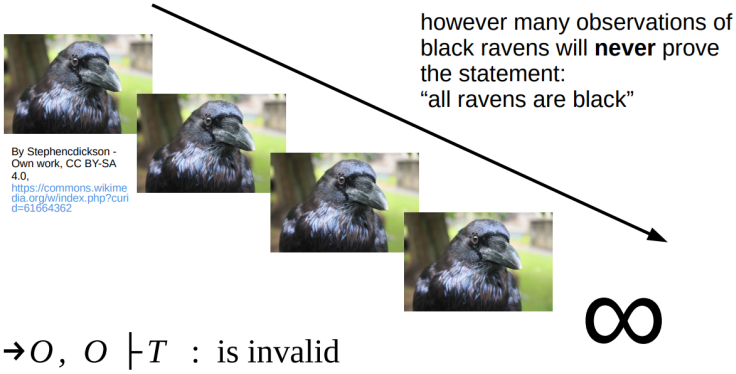
- Scientific statements are characterized by there being procedures that could show the statements to be false
- For non-scientific statements, there are no such procedures

Learning goals:

1. Understanding the induction problem
2. Understanding how falsification follows deductive procedures
3. Understanding how Popper understands falsifiability as a characteristic that separates scientific from non-scientific statements (solves the demarcation problem)

Induction Problem

- Induction: passing from singular statements (single ravens) to universal statements (all ravens)
- We cannot prove by induction, i.e. truth of conclusion doesn't follow from truth of premises
  - a weak version may be that we provide evidence
- Seemingly irrelevant evidence must be admitted as evidence (raven paradox)



$T \rightarrow O,\ O \nvdash T$: is invalid

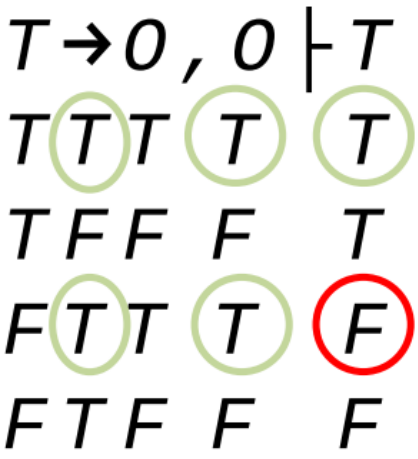
a single observation of a white raven will however disprove the statement: "all ravens are black"

$T \rightarrow O,\ \neg O \vdash \neg T$: is valid

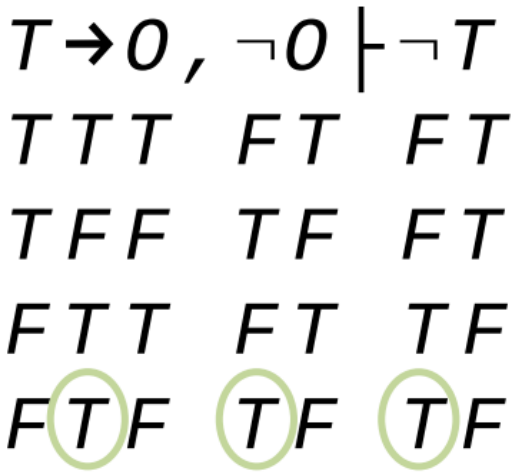
BUT WHY

- i will tell you

Why is it invalid?



Why is it valid?



Principle of demarcation

> It must be possible for an empirical scientific system to be refuted by experience"
> (Popper 2002, p. 18)
> I.e. it has to be falsifiable

Problem of auxiliary theories

- Theories, T, do not exist in isolation
  - They are supported by auxiliary theories, $[A_1, \ldots, A_n]$

$(A_1 \wedge T) \rightarrow O,\ \neg O \vdash \neg T,$ invalid

$(A_1 \wedge T) \rightarrow O,\ \neg O \vdash \neg(A_1 \wedge T)$ valid, but not what we want

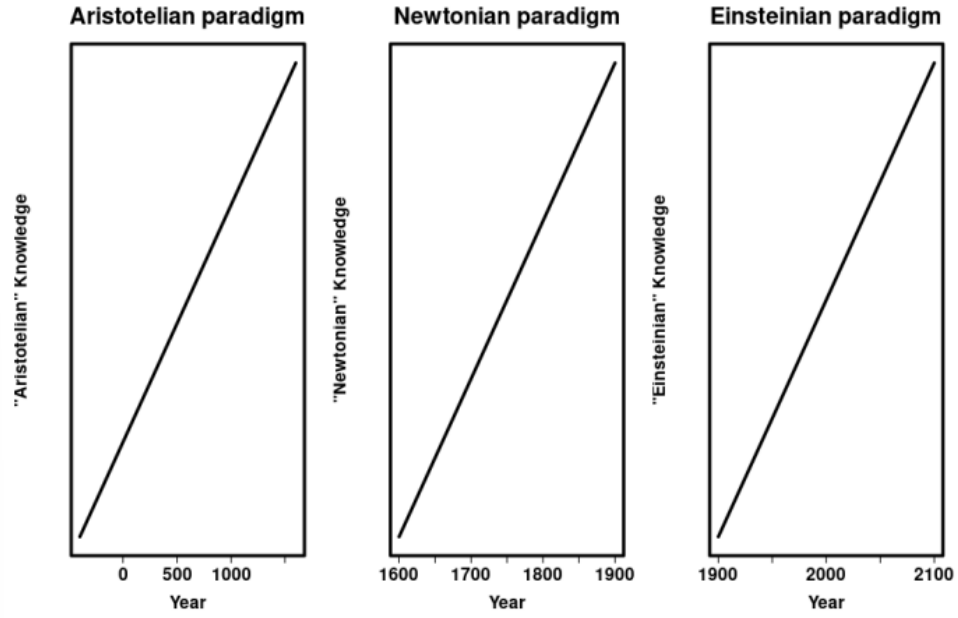
Demarcation of Science: Paradigms

- Progress of science is not linear
- Science is done within paradigms
- Paradigms shape what phenomena can be studied and how these phenomena can be interpreted
- Paradigms are often incommensurable

Learning goals:

1. Understanding Kuhn's idea of paradigms and paradigm shifts
2. Understanding why paradigms may be incommensurable
3. Capability to reflect on what amounts to truth

Knowledge accumulation & incommensurability



Two key insights

1. There is no such thing as neutral observation
2. A theory, T1 , cannot be verified or falsified without considering the grander network of theories, {T1 , T2 , ... , Tn } that T1 is part of

There is no such thing as *neutral* observation

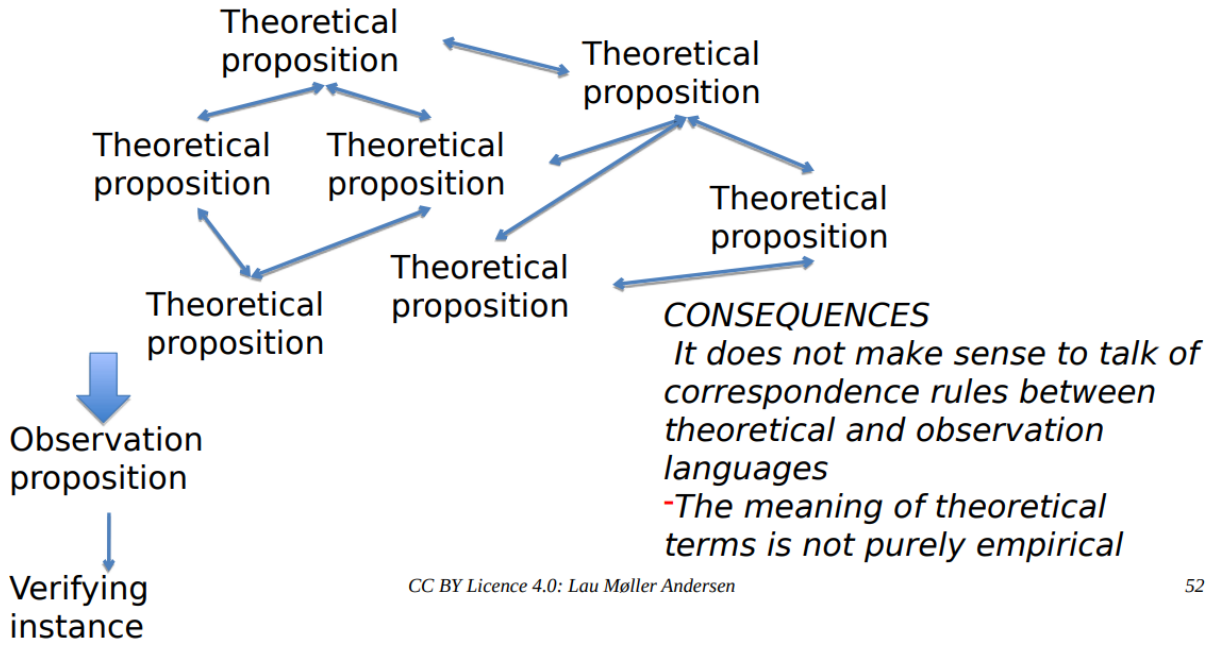
All observation is *embedded* in theory

Rules out the logical empiricist model of theories

..but proposes a holistic model of meaning

Holistic model of meaning

Slide adapted from Joshua Skewes



52

Summary of Philosophy of Science

- Sensation, O, is not neutral, i.e. it depends on our minds (Kant) and our (scientific) theories, T (Kuhn).
  - cf. “the myth of the given”
- Our theories, T, stand in need of translation to entailed observations, O, (Logical positivism).
  - The “correct” translation, however, is dependent on one’s adapted paradigm (Kuhn and Duhem’s thesis)
- Our theories, T, should in principle be refutable (Sober and Popper)
  - in practice, this is highly dependent on one’s auxiliary assumptions, A, (Sober and Quine)

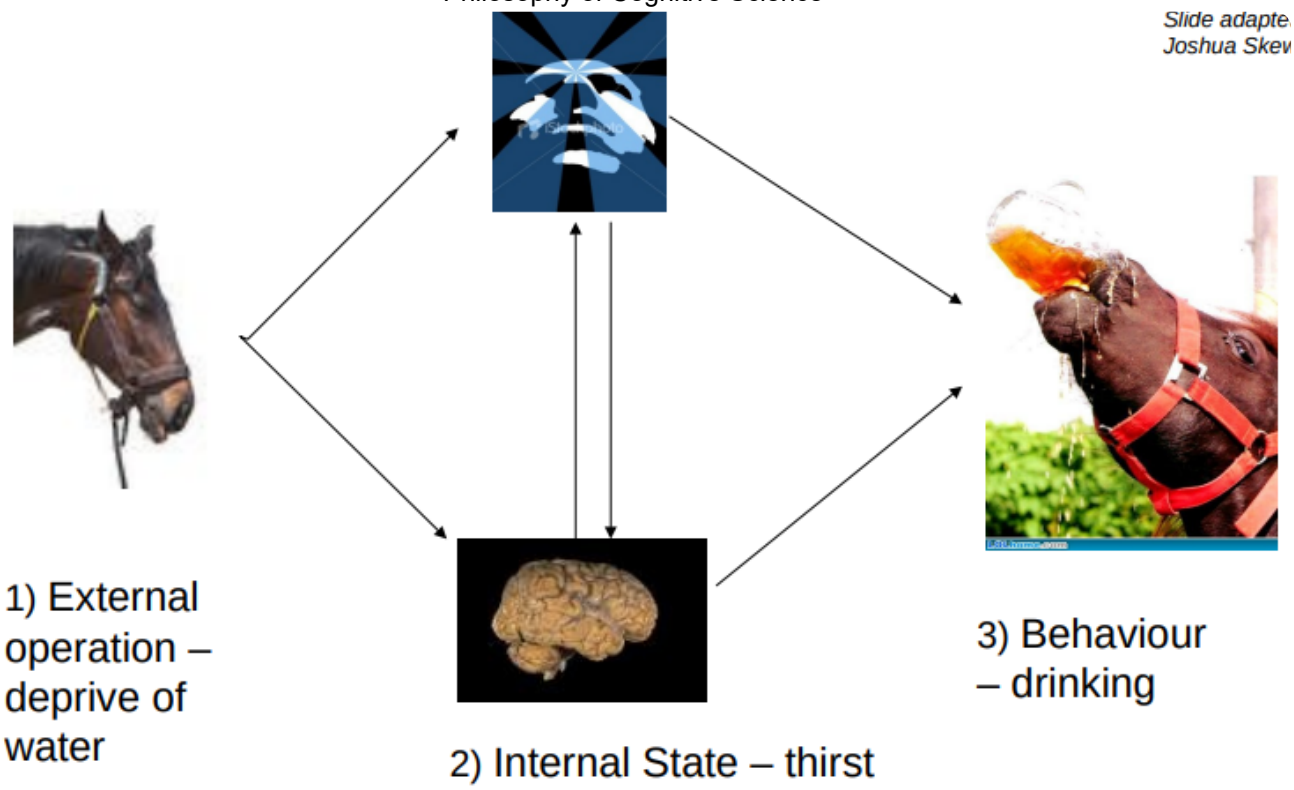
Philosophy of cognitive science

Behaviorism

- The proper study of the mind is the study of behaviour
- Behaviour is operations on the environment
- To predict and control behaviour, one needs to control the environment

Learning goals:

1. Understanding how behaviourism is an externalist paradigm
2. Understanding what operant conditioning is, and how it differs from classical conditioning
3. Understanding the connection to empiricism and logical positivism
4. Being able to critically reflect on the assumptions of behaviourism
5. Being able to see how behaviourism is still relevant today



Two arguments against intermediary variables

First argument

- It is rarely possible to manipulate directly, e.g. “physiological or psychic thirst”
- We can predict behaviour without it, i.e. by looking at the history of operations
- **Thus**: intermediary variables are not *necessary* for predicting and controlling behaviour

Second argument

- Suppose intermediary variables were easy to manipulate directly
- That manipulation would have to be included in an explanation
- That manipulation is just another external operation
- Explaining in terms of intermediary variables would be explanatorily empty without explaining this external operation
- **Thus**: intermediary variables are not sufficient for predicting and controlling behaviour

Against inner causes

- Premise 1: Intermediary variables are not necessary to scientific psychology
- Premise 2: Intermediary variables are not sufficient for scientific psychology
- Conclusion: Intermediary variables have no place in scientific psychology

1. Observe a behaviour
2. Explain in terms of the history of operations
3. Perform further operations and predict outcomes
4. Perform those operations that lead to desired behaviour
5. Create a utopia with a technology of behaviour

- What is an operation?
  - Reinforcement and punishment
- What is a history of operations?
  - Reinforcement schedule
- What is a technology of behaviour
  - A global Skinner box

Critique - final argument

- Premise 0: Intermediary variables are necessarily question-begging in psychological theories

Computationalism

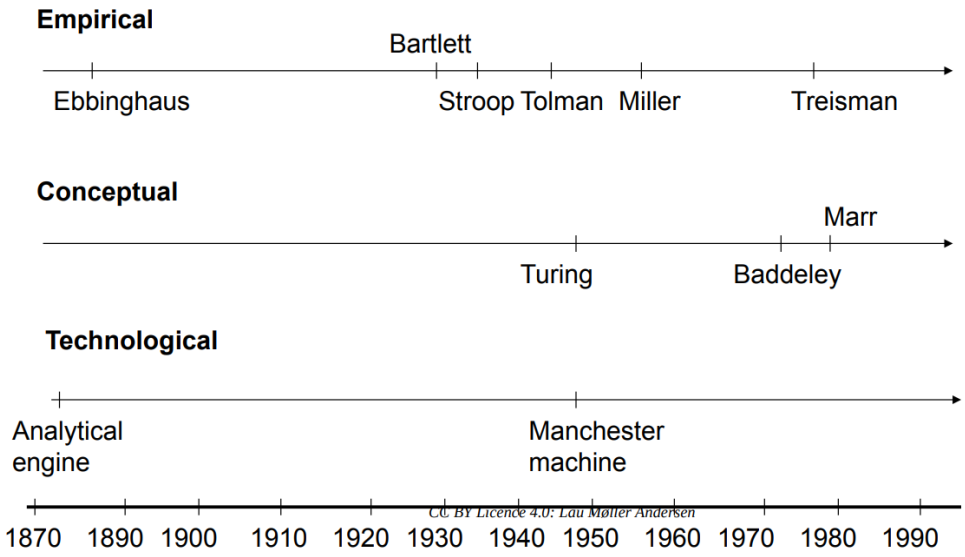
- Cognition is computation on representations
- Computation is implementation independent (it does not what hardware it is run on, e.g. brains or computers)
- Computation is systematically interpretable symbol manipulation (it has strict syntax)

Learning goals:

1. Understanding the intellectual history leading to the cognitive revolution
2. Understanding the assumptions of computationalism
3. Understanding what (mental) representations are
4. Understanding Marr - Vision, A Computational Investigation > The Three Levels

TENSIONS (not an exhaustive list)

TENSIONS (not an exhaustive list)



The computationalist claim

- Human minds work the same way as Turing machines, i.e.
  - Cognition is computation also called C=C
- Computation is systematically interpretable symbol manipulation (it has strict syntax)
- Computation is implementation independent (it is multiply realizable, e.g. by mechanics, computers and brains)

Computationalism and psychology

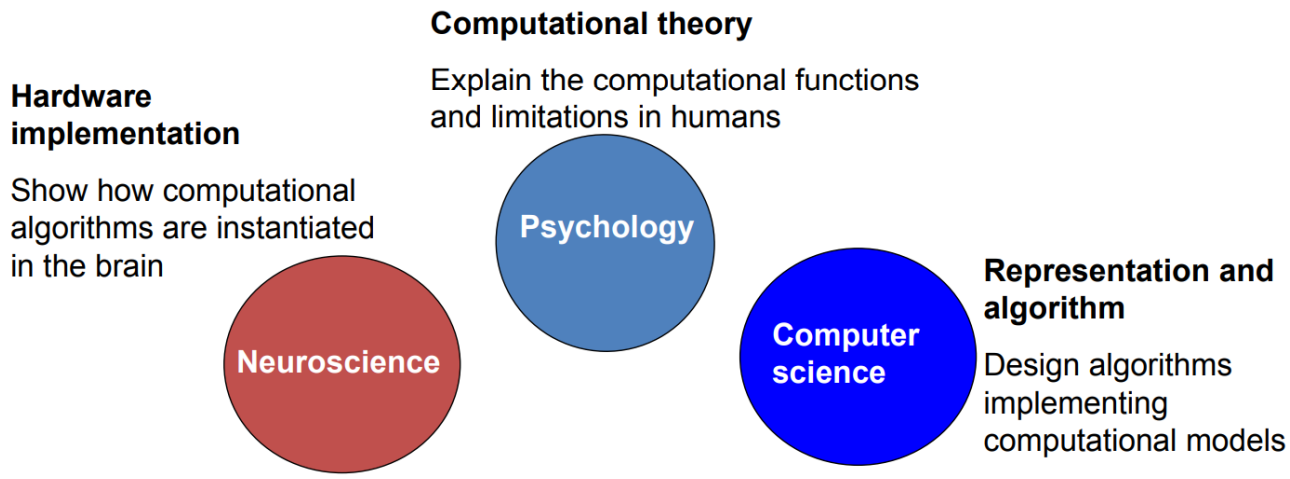
- Under behaviourism
  - Psychology is a science of behaviour
  - Our behaviour is determined by schedules of reinforcement
- Under computationalism
  - Psychology is a science of (human) computation

- Our psychology can be implemented in Turing Machines

Computational theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Figure 1–4. The three levels at which any machine carrying out an information-processing task must be understood.

and



Rough translations to scientific practices

- Psychology = Experiments on function
- Computer Science = Modelling
- Neuroscience = Experiments on tissue

Now we have a programme: do (cognitive) science according to the three levels and reveal facts about (human) cognition

- Cognition is computation cognition is ⇒ implementation-independent, systematically interpretable symbol manipulation
- And a way to assess our models Turing ⇒ indistinguishability

Critique of computationalism

- Computation is not sufficient for cognition

- Syntax does not give you semantics (the so-called grounding problem)
- Grounding problem: how do arbitrary symbols get to stand for something?

Learning goals:

1. Understanding the impact of the Chinese Room Argument
2. Understanding the Symbol Grounding Problem
3. Understanding the different levels of Turing indistinguishability
4. Ability to critically reflect on the scope of the Chinese Room argument

Searle's axioms

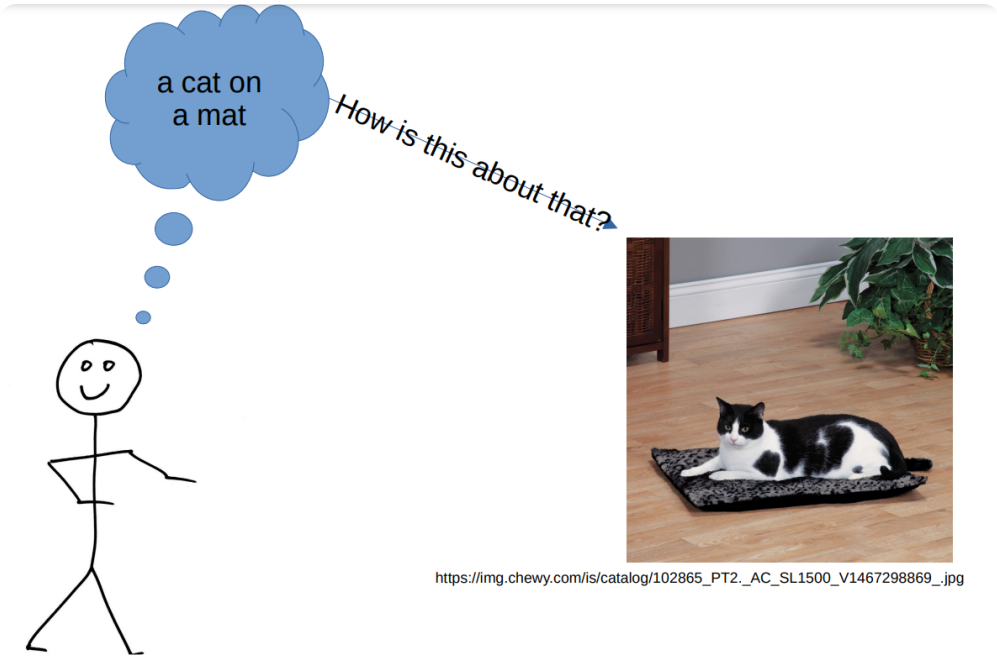
"Axiom 1. Computer programs are formal (syntactic)."
"Axiom 2. Human minds have mental contents (semantics)."
"Axiom 3. Syntax by itself is neither constitutive of nor sufficient for semantics."

Searle's conclusion

"... Programs are neither constitutive of nor sufficient for minds."

Thus: C≠C

> "The symbol grounding problem is [...] that of connecting symbols to what they are about without the mediation of an external interpretation"
> Harnad - Computation is just interpretable symbol manipulation; cognition isn't (p. 386)



https://img.chewy.com/is/catalog/102865_PT2._AC_SL1500_V1467298869_.jpg

The upshot of the symbol grounding problem

- Insisting on *implementation independence*, means that you can only design T2-indistinguishable systems
  - which will not have intrinsic semantics and thus not reflect cognition, which seems to be about something
- To get semantics, you need at least T3-indistinguishability
  - losing implementation independence
- The Chinese room argument does not work against Turing indistinguishability as such
  - only against limited functional mimicry (T2)

see Critique of Computationalism - Lecture > Turing indistinguishability

Connectionism

- Computation happens below the representational level in a connected network of units (e.g. neurons)
- Semantics emerges from the network
- The algorithm level is not independent from the implementation level

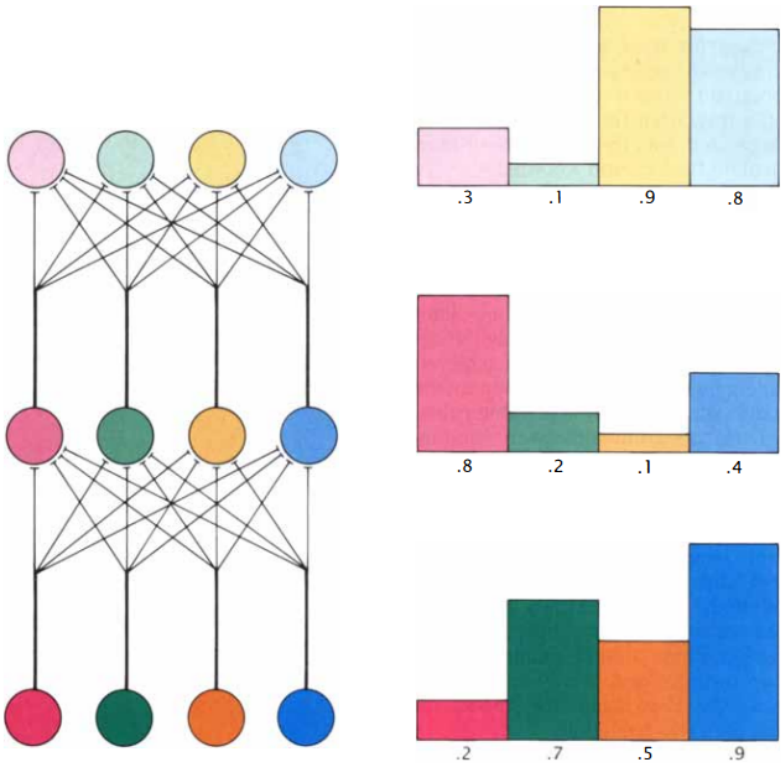
Learning goals

1. Understanding the limitations of the computationalist paradigm seen from the connectionist paradigm
2. Being able to reflect on the values of thought experiments
3. Understanding how neural networks work on a broad level
4. Understanding how semantic content may emerge from the network, i.e. how symbols may get grounded

3 reasons against the luminous room argument and from analogy also against the Chinese room argument

1. Axiom 3 begs the question
2. Our intuitions do not constrain nature
3. At the end of the day, it is an empirical question (whether syntax can cause semantics)

Connectionist networks

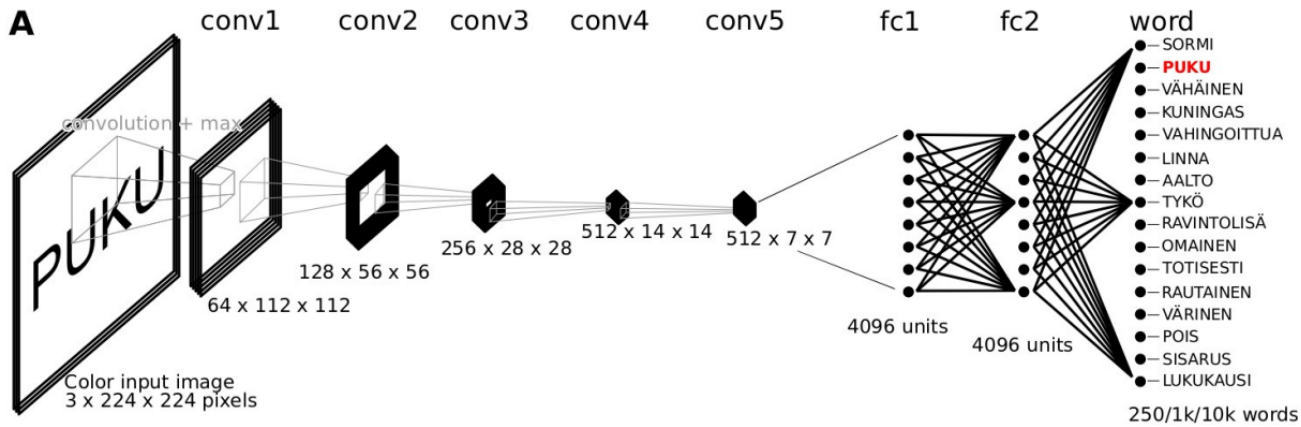


NEURAL NETWORKS model a central feature of the brain's microstructure. In this three-layer net, input neurons (*bottom left*) process a pattern of activations (*bottom right*) and pass it along weighted connections to a hidden layer. Elements in the hidden layer sum their many inputs to produce a new pattern of activations. This is passed to the output layer, which performs a further transformation. Overall the network transforms any input pattern into a corresponding output pattern as dictated by the arrangement and strength of the many connections between neurons.

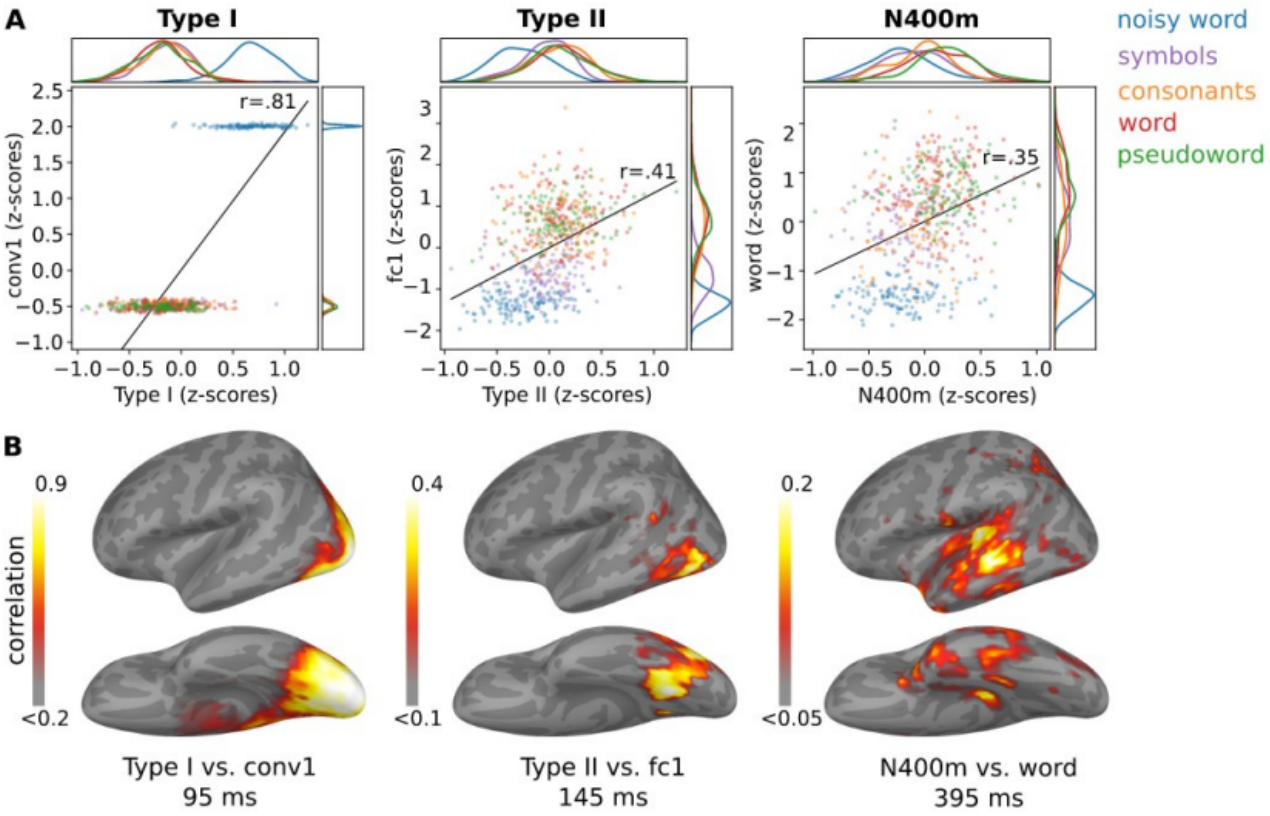
1. Brains are parallel machines
2. Neurons, its processing units are analogue (in terms of spiking frequency)
3. Projections are both feedforward and recurrent

ADVANTAGES OF PARALLELISM

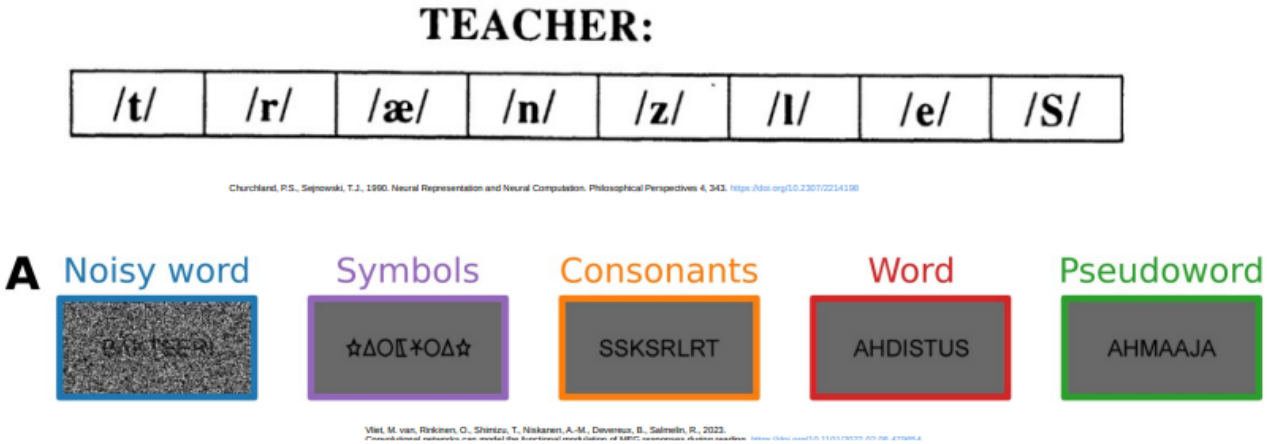
Modern example:



and



External interpretation
CONNECTIONIST NETWORK GOALS ARE SET EXTERNALLY



What if we equip our connectionist networks with sensorimotor processing (T3)?
DO THEY GET CLOSER TO GROUNDING THEIR REPRESENTATIONS?

Robot dog vid - Boston dynamics

Predictive processing

- Perception is causal inference

- Action is active inference
- The mind does prediction error minimization

Learning goals:

1. Understanding how perception can be seen as inference based on sensory input
    - Perceptual inference is done following Bayes' rule
2. Understanding how action is necessary to select the best inference
    - Best inference = the inference that minimizes prediction error
3. The capability to reflect on the relevance of this for the symbol grounding problem and artificial intelligence
4. Appreciating how prediction error minimization may explain mind attributes such as emotion, introspection, privileged access and self

Main assumption of predictive processing

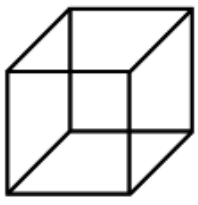
> !! "a [...] substantial view based on the rather uncontroversial idea that the brain is involved in information processing, **and that information theory is cast in terms of the probability theory from which Bayes' rule is derived**" (my emphasis) p. 24

How are likelihood and posterior probability related?

$$P(H|O) \propto P(O|H)\ P(H)$$

Hohwy writes: "inference is a normative notion" (p. 14). What does he mean by that?

Building intuition for "perception is inference":



By Cecilia Bleasdale - https://web.archive.org/web/20150227014959/http://swiked.tumblr.com/post/112073818575/guys-please-help-me-is-this-dress-white-and, Fair use, https://en.wikipedia.org/w/index.php?curid=69200610

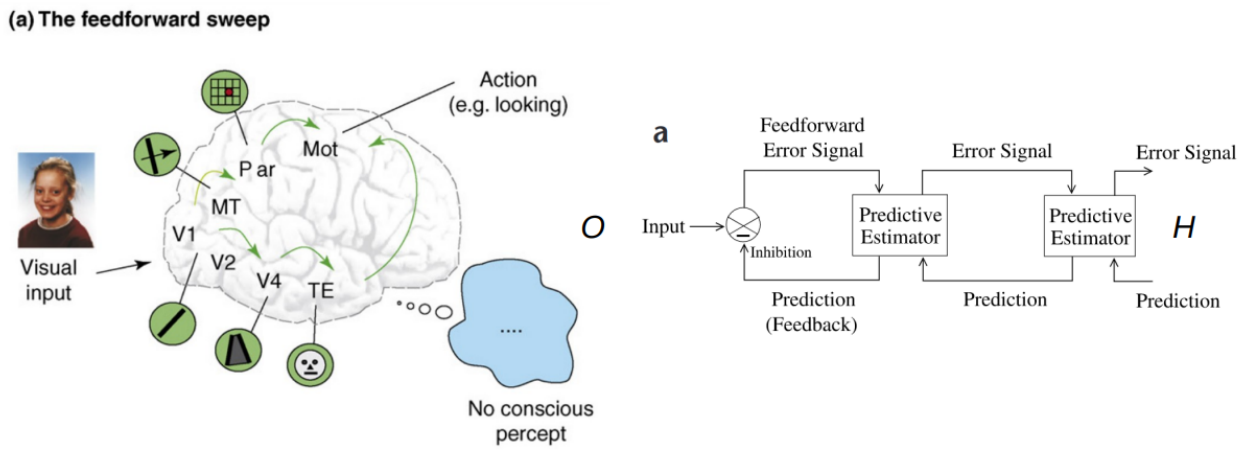
By Fibonacci - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1788215

Letting go of metaphors?

- Computationalism ⇒ the mind as a computer
- Connectionism ⇒ the mind as brain
- Dynamicism/embodiment ⇒ the mind as a Watt Governor
  - just for reference: Van Gelder, T., 1995. What Might Cognition Be, If Not Computation? The Journal of Philosophy 92, 345–381. https://doi.org/10.2307/2941061
- Predictive processing: Explore the mind, by exploring the brain

An inversion of the classical conception

visual cortex example



Symbol grounding solution?

> “The world is the truth. [...] Perhaps the mistake, in earlier conceptions of the problem of perception, was to look for a supervisor (a programmer or the system itself), which somehow knows the truth [...]. The prediction error minimization approach cuts out the middleman and lets the supervisor be the truth itself.”
> (original emphasis) p.49, chapter 2

Consciousness

- Two important theories of consciousness in cognitive neuroscience will be pitted against one another
  - Global workspace theory - Dehaene

    - Conscious items are those we have access to?
- **Recurrent processing theory** Lamme
    - Conscious items are the ones we are phenomenally aware of

Learning goals:

1. Understanding the difference between access consciousness and phenomenal consciousness
2. Understanding, how we may able to investigate consciousness in cognitive science
3. Capability to reflect on how consciousness is seen from the viewpoints of behaviourism, computationalism, connectionism and predictive processing

Are there (conscious) minds

Behaviourism

- Maybe, but only behaviour should be studied and “mind-talk” should thus be translated into behaviour to be studied

Computationalism

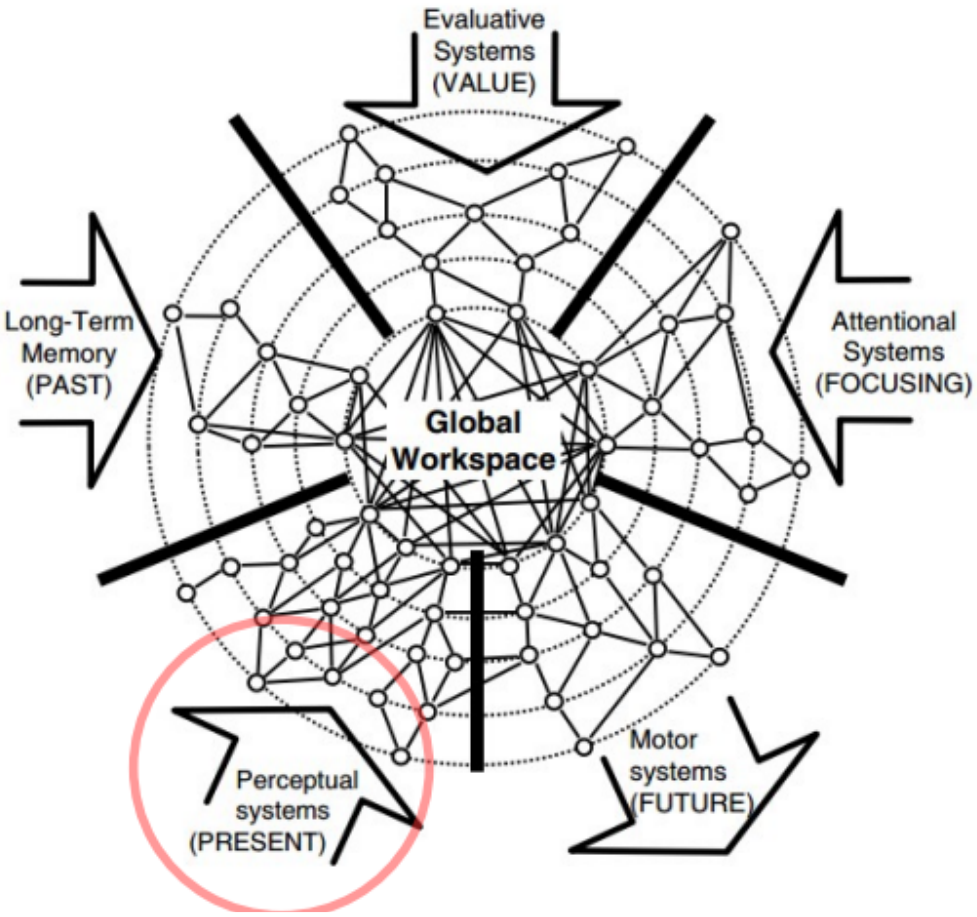
- Yes, and they can be implemented anywhere

Connectionism

- Yes, and for what we know now they depend on brain-like structures

Predictive processing

- Yes, and they are dependent on brains that build models of the world (and update them based on prediction errors)



l. 2014

Two kinds of consciousness?

- Phenomenal
  - Related to the content (semantics)
  - Related to the ‘feel’
  - Does not fit an input output model
- Access
  - Availability to other systems, e.g. reasoning
  - Fits an input-output model

Paradigm issues

Reprise - are there conscious minds?

Behaviourism

- Translate mental words into behaviour: X desires p, e.g.: If X is in the vicinity of p, X will try to obtain p
- **Issues**: How do we operationalize conscious experience? (I.e. how would it manifest in behaviour?)

Computationalism

- Da Three Levels.png
- **Issues**: Is conscious experience information processing? Does it fit into an input-output model? Why would there be something it is like to process information?

Connectionism

- Build brain-like networks that can emulate mindful behaviour
- **Issue**: Even if we accept that semantics emerge from the distributed network as high-dimensional vectors, how and why would a brain-like model instantiate conscious experience? (symbol grounding problem)

Predictive processing

- Model the mind as a hierarchical structure doing information processing e.g.
    - Interpret perception as providing prediction errors for higher-order models of what the world is like; and allow for action, active inference, to update the models by providing new errors
- **Issue**: Why is subjective experience bound to arise from hierarchical information processing, even if these models are created from a first-person perspective?

The crux of the matter:

- All our paradigms are functional* , i.e. they aim to explain how and why goal-directed behaviour emerges
- However, what is the function of phenomenal consciousness?
        - might predictive processing be an exception?

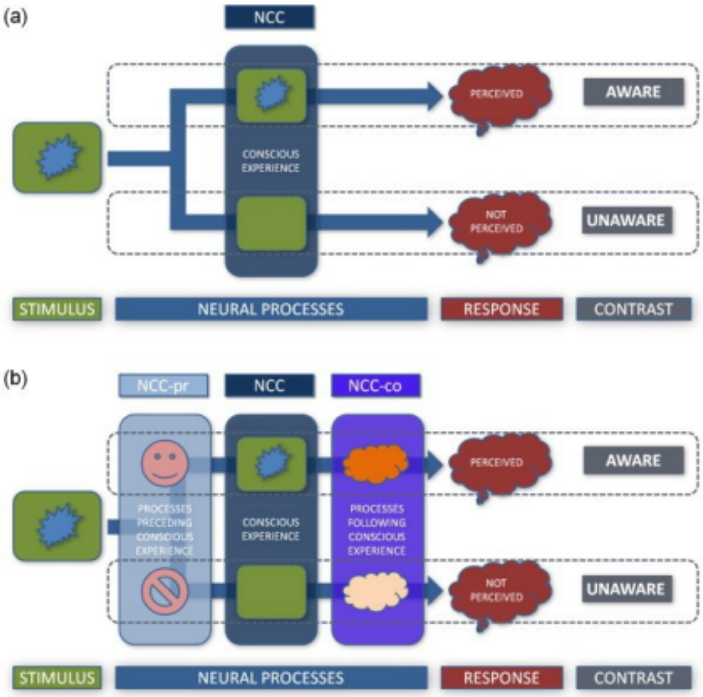
If there is no function...

- Consciousness is just an **epiphenomenon**
    - Epiphenomenon: A secondary phenomenon is one that occurs alongside a primary phenomenon (the mind's functional workings), but has no causal connection to world, i.e. conscious experience is a by-product of the mind

> Leaving aside the function, can we, at least, find correlates of phenomenal consciousness using cognitive science?

Contrastive analysis – take II

too simplistic?



Summary of Philosophy of Cognitive Science

- Behaviourism can be seen as the proto-paradigm of cognitive science
  - keep Morgan's canon in mind – don't use cognitive explanations if behavioral will do
- Cognitive science proper begins with computationalism
  - The aim of both computationalism and connectionism is to emulate human cognition using metaphors of the mind
  - This is often done using artificial intelligence
  - Predictive processing claims to break out of the metaphors and just model the mind
- The latter three paradigms try to come to terms with how to solve the symbol grounding problems, i.e. how do representations get to be about something
  - This is related to consciousness, why is there something it is like to have a representational mind
- These paradigms shape how we think, and what we consider worthwhile problems
  - and they do not seem to be falsified in one fell swoop

Justified True Belief

#Philosophy

Justified true belief = Knowledge

Justified

- Epistemology = How we should acquire knowledge

Truth

- Ontology = The nature of being

Belief

- Cognitive science - what makes us form beliefs
- What can we know?
  - Skepticism
- How do we know?
  - Sensation
  - Reason
- What kinds of knowing
  - analytic, synthetic, *a priori*, *a posteriori*
  - manifest image vs scientific image

Descartes

Logical Connectives (Wiki)

#Philosophy

Taken from: Logical Connective
and: List of Logic Symbols
as well as now: LaTeX Cheatsheet

> !! In formal languages, truth functions are represented by unambiguous symbols. This allows logical statements to not be understood in an ambiguous way. These symbols are called logical connectives, logical operators, propositional operators, or, in classical logic, truth-functional connectives. For the rules which allow new well-formed formulas to be

constructed by joining other well-formed formulas using truth-functional connectives

Symbol, name	Truth table	Venn diagram
Zeroary connectives (constants)		
⊤ Truth/tautology	1	
⊥ Falsity/contradiction	0	
Unary connectives		
	p = 0 1	
Proposition p	0 1	
¬ Negation	1 0	
Binary connectives		
	p = 0 1 q = 0 1 0 1	
Proposition p	0 0 1 1	
Proposition q	0 1 0 1	
∧ Conjunction	0 0 0 1	
↑ Alternative denial	1 1 1 0	
∨ Disjunction	0 1 1 1	
↓ Joint denial	1 0 0 0	
→ Material conditional	1 1 0 1	
↮ Exclusive or	0 1 1 0	
↔ Biconditional	1 0 0 1	
← Converse implication	1 0 1 1	

More information

List of Common Logical Connectives

- **Negation (not)**: $\neg$, $\sim$, N (prefix) in which $\neg$ is the most modern and widely used (as well as $\sim$)
- **Conjunction (and)**: $\wedge$, $\&$, K (prefix) in which $\wedge$ is the most modern and widely used
- **Disjunction (or)**: $\vee$, A (prefix) in which $\vee$ is the most modern and widely used
- **Implication (if...then)**: $\rightarrow$, $\supset$, $\Rightarrow$, C (prefix) in which $\rightarrow$ is the most modern and widely used and $\supset$ is used by many too
- **Equivalence (if and only if)**: $\leftrightarrow$, $\Leftrightarrow$, $\equiv$, E (prefix) in which $\leftrightarrow$ is the most modern and widely used

For example, the meaning of the statements *it is raining* (denoted by p) and *I am indoors* (denoted by q) is transformed, when the two are combined with logical connectives:

- It is *not* raining ($\neg p$)
- It is raining *and* I am indoors ($p \wedge q$)
- It is raining *or* I am indoors ($p \vee q$)
- *If* it is raining, *then* I am indoors ($p \rightarrow q$)
- *If* I am indoors, *then* it is raining ($q \rightarrow p$)
- I am indoors *if and only if* it is raining ($p \leftrightarrow q$)

It is also common to consider the *always true* formula and the *always false* formula to be connective

- **True** formula: $\top$, 1, V (prefix), or T
- **False** formula: $\bot$, 0, O (prefix), or F

English word	Connective	Symbol	Logical gate
not	negation	$\neg$	NOT
and	conjunction	$\wedge$	AND
or	disjunction	$\vee$	OR
if...then	material implication	$\rightarrow$	IMPLY
...if	converse implication	$\leftarrow$	
either...or	exclusive disjunction	$\oplus$	XOR
if and only if	biconditional	$\leftrightarrow$	XNOR
not both	alternative denial	$\uparrow$	NAND
neither...nor	joint denial	$\downarrow$	NOR
but not	material nonimplication	$\nrightarrow$	NIMPLY

- ALSO: Therefore = $\therefore$

- ALSO: Syntactically entails = $\vdash$

Propositional Logic

Modus Tollens (denying the consequent)

- a mixed hypothetical syllogism that takes the form of "If P, then Q. Not Q. Therefore, not P."
  - $P \rightarrow Q, \neg Q \therefore \neg P$
-

Mock Exam

Discuss how the views of science differ between Kuhn and Popper - how do their views apply to cognitive science and what are the substantial differences between their views

Kuhn's view of science

- Paradigm shifts
  - In Kuhn's view science comes in paradigms, where one paradigm is replaced by another after a scientific revolution
  - He argued that science progresses through periods of normal science, where a dominant paradigm guides research, and revolutionary science, where anomalies and crises lead to a shift in fundamental beliefs.
- Incommensurability
  - Kuhn suggested that different paradigms are often incommensurable, meaning that they use different concepts, methodologies, and standards of evaluation. This makes it difficult to compare or judge them objectively.
- Historical context
  - Kuhn emphasized the importance of historical and social factors in shaping scientific practice. He argued that scientific theories are influenced by the socio-cultural context and that scientific communities operate within shared worldviews.
    - Scientists are often deeply invested in the prevailing paradigm, and resistance to change can be strong.

Popper's view of science

- Popper criticized induction (the induction problem) and logical positivism (verifiability)
- Falsifiability
  - Science must be falsifiable (contrary to the logical positivists; verifiability criterion)
  - Popper focused on the criterion of falsifiability as the demarcation between scientific and non-scientific theories. According to him, a scientific theory must be falsifiable, meaning there must be conceivable observations that could refute it.
  - You can never prove a theory in Popper's view but only corroborate it (strengthen its claims)
- Demands objectivity

Application to Cognitive Science

Kuhn in Cognitive Science: Kuhn's ideas can be applied to cognitive science by recognizing shifts in dominant paradigms. For example, the transition from behaviorism to cognitive psychology can be viewed as a paradigm shift.

- Their achievements were sufficiently unprecedented which made their paradigm convincing enough $\rightarrow$ dominates in NA

Popper in Cognitive Science: Popper's falsifiability criterion is relevant in evaluating cognitive theories. The ability to subject cognitive models to empirical testing and potential falsification is crucial for considering them scientifically valid.

- It gets tricky though when you consider the Duhem-Quine thesis (that all theories are reliant on auxiliary theories) which is especially the case inside of cognitive science where most of our hypotheses rest on a foundation of many many different theories of interlinked structures in cognition

Differences in Approach: Kuhn's emphasis on the role of historical and social factors might lead to a more relativistic view in cognitive science, acknowledging that different research communities might operate under distinct paradigms. Popper's emphasis on falsifiability and objective evaluation may lead to a more universal and stringent approach to theory assessment.

Philosophical statements

This Reddit thread does a great job of explaining the distinctions between the categories as well as touch upon Kant's vs Hume's epistemology

On the difference between philosophical statements

- **Analytic**
  - true by virtue of meaning
- **Synthetic**
  - true based on empirical evidence
- *a priori*
  - known independently of experience
- *a posteriori*
  - rely on experience for their verification

	Meaning	**Example sentences**	**Associations**
Analytic	truth or falsehood can be determined *solely* by analyzing the meanings of the terms involved, without reference to any external facts or empirical evidence	"All bachelors are unmarried men" the concept of bachelor *already includes* the idea of being unmarried	closely associated with logic and are considered necessarily true.
Synthetic	truth or falsehood depends on *empirical* facts about the world, and they cannot be determined solely by analyzing the meanings of the terms involved	"The cat is on the mat" describes a state of affairs in the world that needs to be *verified through observation*.	the basis of scientific inquiry and are contingent upon empirical evidence.
a priori	can be known to be true or false *independently* of experience or empirical evidence	"2 + 2 = 4" is a priori because it can be understood and proven true using *logical reasoning alone*	often associated with necessary truths and are considered to be *true in all possible worlds*

	Meaning	Example sentences	Associations
a posteriori	statements whose truth or falsehood can *only be known through* experience or empirical evidence	"Water boils at 100 degrees Celsius at sea level" is a posteriori because it can *only* be verified through empirical testing.	*contingent* upon the particular facts of the world and are not necessarily true in all possible worlds.

Now Kant, Critique of Pure Reason > Introduction (as in the second edition) Sections I-V introduces a mixture of these types of statements and talks about their role in epistemology

And a bit more in depth: (from the above Reddit thread)

A Priori - *A Posteriori* refers to the method of verification of the judgement.

- An *a priori statement* is one who's truth can be assessed without having to rely on any experience
- an *a posteriori* statement's truth can only be verified by reference to a specific experience.
- So this distinction refers to the relationship between the world and the judgement uttered.

Synthetic - Analytic refers to the inner structure of the statement, more specifically, to the relationship between Subject and Predicate.

- A basic judgement takes the form *S is P* where S and P are nouns or noun phrases.
- Now, for analytic judgements, the standard formulation ("*Container formulation*") is this:
  - A judgement is analytic if the predicate term is "*contained*" in the subject term.
  - For example, the concept of the "unmarried man" is contained in the concept of the "bachelor".
  - A synthetic judgement is then one in which subject and predicate are, in themselves, unrelated.
  - For example, I could say that this ball in front of me was red. Naturally, the concept of redness is not "contained" in the concept of the ball, I could just

as well imagine blue and yellow ones.

This is the part where my preliminary characterization in the first part of the last paragraph breaks down because for Kant, a synthetic judgement is one in which two concepts are united.

- How does this unity come to be?
  - By reference to something third.
- This third thing that unites the concepts "ball" and "red" and which, we might say, corresponds to the copula "is" in the judgement, is the actual ball right here in front of me that is given in my experience.

This is where Kant differs from the empiricists: The empiricists claim that in such a judgment, two ideas are simply "associated" with each other, simply "held next to each other". Kant is saying that this won't do. A synthetic judgement for Kant is one which "picks out" a specific object in the real world for to which the two concepts apply.

Now you can mix and match these two categories to get 4 different kinds of statements:

1. Analytic a priori
2. Analytic a posteriori
3. Synthetic a posteriori
4. Synthetic a priori

Category 1 are the standard analytic statements. These are basically just definitions. Things like "All bachelors are unmarried men" simply define what it is to be a bachelor.

For Kant, category 2 is basically *superfluous* because if a Judgement is analytic, I don't *need* to "look out into the real world" to verify it. I can just **analyze** the two concepts and see if one is implied by the other.

Category 3 are the standard empirical judgements we make on a day to day basis, like "This drink tastes bitter". It's synthetic because "tasting bitter" is not necessarily contained in the concept of a drink and a posteriori because it references an object "out there in the world"

But really, why all this fuss about judgements? Simply put, Kant associates the *crucial* concepts of **universality** and **necessity** with the *a priori*. Anything experience teaches me is particular and *contingent*, it could be different.

- I may observe an apple falling from a tree but I could just as well imagine the apple to fly upwards.

For Hume, there were only two types of judgments: "matters of fact" (synthetic a posteriori) and "relations of ideas" (analytic a priori).

- Thus, for example, if you wanted to prove that certain causal laws applied to the real world, you'd have to ask yourself, what kind of judgements am I uttering here?
- Is it a "relation of ideas" that apples fall down? No, analyze the two concepts as much as you want, it doesn't simply from definitions that apples fall.
- Can I verify that claim through experience?
  - *Certainly!* **BUT** this verification would only be a *particular* one. It would only refer to this particular apple here in front of me.

Can I know that the world adheres to certain necessary laws? For Hume, this statement would imply that the future behaves just like the past.

- Now, ask yourself, what kind of a judgement is that? Is it a relation of ideas? No.
- Is it a matter of fact? Clearly not since nobody has had an experience of the future.
- So for Hume, universal and necessary laws of nature cannot be demonstrated to exist.

This is really bad. (lol)

Now, Kant claims that Hume went wrong by not realizing that there is a third kind of judgement, the famous *synthetic a priori*

- These are judgments in which something substantial is added to the subject but which are nevertheless knowable without recourse to experience.
- Crucially, this means that they are universal and necessary.
  - Statements like "the shortest path between two points is a line" or "bodies are heavy" belong to this category.
  - These are the statements mathematics and natural science makes. Kant assumes that Newtonian science is universally and necessarily true, this for him is just an obvious given.
  - His analysis shows that the claims of science are, at heart, synthetic a priori.

Now, the big question: How are synthetic a priori statements possible?

- If we can answer that, then we will have explained the validity of the sciences beyond any shadow of a doubt and will have done away with that insane Humean skepticism.

Now, the short answer is that synthetic a priori judgements are possible by being those judgements that make experience possible and without which experience could not be conceived of.

- This is the task of the rest of the Critique of Pure Reason.

Truth Tables

see also Logic

Modus ponens
$R \rightarrow B, R \vdash B$

R	→	B	,	R	⊢	B
T	T	T		T		T
T	F	F		T		F
F	T	T		F		T
F	T	F		F		F
All Ravens	are	Black		It's a Raven	so	it's Black

= Valid because there's truth preservation

Modus ponens but theory/observation

T: a theory
O: an observation
$T \rightarrow O$: T entails O

$T \rightarrow O,\ O \vdash T$: is invalid

$T \rightarrow O, \neg O \vdash \neg T$: is valid

Induction problem
$R \rightarrow B, B \nvdash R$

R	→	B	,	B	⊬	R
T	T	T		T		T
T	F	F		F		T
F	T	T*		T		F
F	T	F		F		F
All Ravens	are	Black		It's a Black	so	it's a Raven

*A bird can be black without being a raven. A street can be wet without it raining

= Invalid because there are two rows with true premises but the conclusions are both T and F (see row 1&3)

Modus tollens
$R \rightarrow B, \neg B \vdash \neg R$

(Illustrates Poppers falsification through the asymmetry)

R	→	B	,	¬	B	⊢	¬	R
T	T	T		F	T		F	T
T	F	F		T	F		F	T
F	T	T		F	T		T	F
F	T	F		T	F		T	F
All Ravens	are	Black		It's not	Black	so	it's not	a Raven

Valid because the premises are only valid once, and there the conclusion is also valid = truth preservation

Something invalid
$R \rightarrow B, \neg R \nvdash \neg B$

R	→	B	,	¬	R	⊬	¬	B
T	T	T		F	T		F	T
T	F	F		F	T		T	F
F	T	T		T	F		F	T
F	T	F		T	F		T	F
All Ravens	are	Black		It's not	a Raven	so	it's not	Black

Invalid because there are two rows with true premises but the conclusions are both T and F (see row 3&4)

T	→	O	,	T	⊢	O
T	T	T		T		T

T	→	**O**	,	**T**	⊢	**O**
T	F	F		T		F
F	T	T		F		T
F	T	F		F		F

Induction Problem (Verification)

T	→	**O**	,	**O**	⊢	**T**	
T	T	T		T		T	
T	F	F		F		T	
F	T	T		T		F	
F	T	F		F		F	

T	→	**O**	,	¬**T**	⊢	¬**O**
T	T	T		F T		F T
T	F	F		F T		T F
F	T	T		T F		F T
F	T	F		T F		T F

Falsification (Popper's argument)

T	→	**O**	,	¬**O**	⊢	¬**T**
T	T	T		F T		F T
T	F	F		T F		F T

T	$\rightarrow$	**O**	**,**	$\neg$**O**	$\vdash$	$\neg$**T**
F	T	T		F T		T F
F	T	F		T F		T F

Philosophers

Empiricism:

- Hume

Kant would be in the middle or *in between* empiricism and rationalism

- Synthetic/a priori: Two straight lines cannot create an enclosed figure

Rationalism:

- Popper (but still somewhere in the middle)