Being a Zombie

Learning, Representation, and Consciousness

Logistics

- Web site is up and running:
  - http://srsc.ulb.ac.be/AI/
What do the following situations have in common?

Athletic skills ...
Learning language ...

bleh
bleh
BLAH!
bleh bleh bleh

bleh bleh bleh
Implicit knowledge!

Many of the things we learn to do:
* are learned without intention,
* without verbalizable knowledge of what was learned,
* and sometimes without knowledge that we learned anything

**IMPLICIT LEARNING:**
A change in performance that is not accompanied by a corresponding change in the ability to describe the acquired knowledge

We learn all the time

- Considerable new evidence for plasticity in the brain:
  - Expert string players exhibit larger-than-normal areas of the somatosensory cortex dedicated to representing input from the fingering digits (Elbert et al., 1995)
  - Posterior hippocampus is enlarged in experienced taxi drivers compared to subjects who do not have extensive experience in memorizing complex maps (Maguire et al., 2000)
  - The very organization of the somatosensory cortex (the famous Penfield homunculus) might depend on pre-natal sensory experience (Farah, 1997)
  - Evidence for neurogenesis was also found in humans, overturning decades of unquestioned — but, as it turns out, erroneous — assumptions about the lack of regenerative cellular processes in the adult brain.
  - Evidence for fetal learning (van Helteren et al., 2000)
  - Evidence for memory consolidation during REM sleep (Maquet et al., 2000)
Plasticity is pervasive and powerful

- Learning is a mandatory consequence of processing
- Experience plays a fundamental role in cognition, not only because adaptation is a fundamental element of successful behavior, but also because experience actually shapes cognition all the way down to the manner in which specific neural circuits process information

Learning in the brain

- Long term potentiation and depression (LTP & LTD)
  - Role of NMDA receptors (sensitive to NT glutamate)
  - Depends on concurrent pre- and post-synaptic activity
- Learning always involves modifying connections
- Distinction between weight-based and activation-based processing
Computational objectives of learning

- Distinction and complementarity between “model” learning and “task” learning (O’Reilly, 1998):

  - **Model learning** involves activity-based hebbian plasticity and is a prime candidate for implicit learning in that it does not require intentions, goals, or error information and is highly sensitive to correlational structure. Conditioning is an example of the operation of such mechanisms.

  - **Task learning** involves error-based plasticity and is a prime candidate for explicit learning in that it is most appropriate for learning specific tasks, that is, learning complex input-output mappings that require specific actions to be related to specific goals.

  - An important goal of current research is to determine exactly how and where in the brain these learning mechanisms combine. One answer is the Leabra algorithm

Model learning

- Basic idea: Capture the statistical structure of the world by associative learning so as to build detailed, accurate, predictive, internal models:

  - Difficult because:
    - What is available through our sense organs is impoverished
    - Inversion is required: Constructing models of the world involves inverting the mapping between the world and our sense organs (an ill-posed problem: multiple models are possible)
Task learning

- Basic idea: Learn the complex input-output mappings required for mastery of specific tasks so as to fulfill our goals

- Requires that the goal behavior be available to the system so that an error-correcting learning procedure may be used
  - **Backpropagation** (generalized delta rule) is the most well-known of such procedures: Hundreds of existing models
  - **GeneRec** (generalized recirculation algorithm) is the most biologically plausible because it does not involve passing around error signals but is instead based on comparisons between unit activity in expectation and outcome phases of processing. This can be achieved using known LTP and LTD mechanisms

- In both cases, learning is based on comparing expectations with actual outcomes: The brain is a prediction machine

Being a zombie...

- Great knot of ultra-complex issues:
  - **Rules vs. associations:**
    - Can connectionist networks learn abstract knowledge?
    - Are associative learning mechanisms sufficient to account for all of cognition?
  - **In what sense can cognition be unconscious?**
    - Can a representation be unconscious?
    - Which representations can be unconscious?
    - How do we find out?
  - **What is the function of consciousness?**
  - **What happens during cognitive development and learning?**
    - Do we start with modules or are modules the result of development?
Controversial issues in many domains

- Language learning and processing:
  - Pinker vs. the connectionists: Rules vs. associations
  - Marcus (1999) experiments with the habituation paradigm
  - Christiansen & Chater on distributional approaches to language acquisition
  - Saffran, Aslin, Gomez, etc.
  - Pacton et al. (2001) on the development of sensitivity to orthographic regularities

- Cognitive Development:
  - Object permanence
  - Knowledge of the rules of physics

- Implicit learning:
  - Transfer experiments in AGL and in sequence learning experiments: Evidence for abstraction?

Abstract knowledge

“Abstraction refers to the generality of production rules. Production rules do not require that a specific stimulus be present; the rules will apply in any stimulus condition that satisfies the pattern specification of the condition” (Anderson, 1981)

“a subject who learned a useful rule would have equal success in transfer on stimuli presented either in the same or different features, because the rule is applicable regardless of the features in which items are presented” (Whittlesea and Dorken, 1997)

- Abstract knowledge is knowledge that is independent from the surface features of the material: Variables, rules specifying which abstract relationships hold, regardless of which exact stimuli are involved
Rule-like vs. rule-based behavior

- Principle 1 — Functional representations:
- Sensitivity to some regularity does not necessarily imply that the regularity itself is represented in the cognitive system as an object of representation
- Distinction between semantically transparent explanatory models vs. emergent explanatory models (Clark, 1990)
- Main point: Rule-like behavior does not entail rule-based representations

Abstraction in networks

“… what the connectionist algorithm does is find correlations between sets of variables that it is given. Things that cannot be expressed in terms of correlations between given variables cannot be found by the algorithm. That is how Steven Pinker at MIT was able to show that you cannot claim that neural nets learn the past tense in English. In principle, they cannot because there are features in the past tense, for example the division of verbs into classes that form their past tense differently that cannot be expressed simply as correlations between variables given to the machine as input” (Putnam, 1995)

- (Erroneous) conclusion: Connectionist networks cannot learn abstract relationships
Marcus (2001)

“According to this view, there is an innately given representational format that allows for the possibility of operations over variables, an innately given set of operations that can be computed over these variables, and innately given apparatus for combining those operations, an innately given representational format that accommodates structured combinations, and an innately given representational format that accommodates representing individuals distinctly from kinds” (Marcus, 2001, pp. 143-144)

Functional similarity

“The relevant overlap of representations required for generalization in a neural network or other statistical learning procedure need not be present directly in the “raw” input but can arise over internal representations that are subject to learning” (McClelland & Plaut, 1999)

- Abstraction as emerging out of the processing of exemplars
Memorize this!

- TSSXS
- TXXVPXVV
- PTTVV
- TXXVVPS
- PTVPXVV
- PTVPXVPS

G or not?

<table>
<thead>
<tr>
<th>String</th>
<th>✔️</th>
<th>✗</th>
</tr>
</thead>
<tbody>
<tr>
<td>TXXVPXVV</td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td>PTTVVV</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>VTVPXVV</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>PVPSS</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>TXXTVVV</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>TVPSTXX</td>
<td>✗</td>
<td></td>
</tr>
</tbody>
</table>

- All the strings you have seen have been generated according to specific grammar rules
- Now you have to decide, for each string, whether it is grammatical or not
Reber (1967) used this grammar in the first artificial grammar learning experiment. The grammar is a simple finite-state automation: Strings of letters are generated by entering the grammar at node S1, moving from node to node and concatenating the labels of the traversed arcs until the end node is reached. Subjects typically do better than chance would predict. Subjects are typically unable to indicate what the rules of the grammar are. Reber’s conclusion:

- People have learned the rules of the grammar unconsciously

**IMPLICIT LEARNING**
Implicit learning

  “The process by which knowledge about the rule-governed complexities of the stimulus environment are acquired independently of conscious attempts to do so.”

- Lewicki (1987):
  “ [...] subjects are able to acquire specific procedural knowledge (i.e. processing rules) not only without being able to articulate what they have learned but even without being aware that they had learned anything.”

- Perruchet & Vinter (1998):
  “The term implicit learning designates an adaptive mode in which subject’s behavior is sensitive to the structural features of an experienced situation, without the adaptation being due to an intentional exploitation of subjects’ explicit knowledge about these features.”

What is learned?

- Rule abstraction approaches produce symbolic knowledge of the material in the form of production rules, discrimination trees, or classifiers.
  “IF the string begins with T or P THEN the string is grammatical.”

- Exemplar-based approaches assume that whole instances are memorized during training. New exemplars can then be classified according to their similarity with either specific items or with the entire memorized database.

- Fragment-based and chunking approaches exploit the redundancy of the training material by decomposing it into short chunks such as bigrams or trigrams. The resulting database can be organized hierarchically or not. New exemplars are classified according to how many chunks they share with the training material.

- Distributional and statistical approaches (including neural network models) develop superpositional representations of the statistical constraints present in the material based on associative learning mechanisms.

Many possibilities exist. Which one is right? Might several be right at the same time? Does the nature of what is learned correlate with availability to conscious awareness?
Rule-based approaches

- Not convincingly substantiated
  - Best (but nevertheless controversial) evidence: Transfer situations in AGL

- Conceptual problem:
  - What is a “rule”?
  - Requires one to assume that we can hold unconscious rules that function just like conscious rules, only minus consciousness

- Methodological problem
  - How do you demonstrate that people have a “rule” when they cannot report it?

Exemplar-based approaches

- Very powerful, but:
  - It is ultimately implausible to assume that our memory stores millions of raw exemplars
  - It is obvious that we need more abstract, structured, compositional representations at some point
Universal learning mechanisms in cognitive architectures such as SOAR

General idea:

- We continuously “parse” the world so as to form increasingly familiar “chunks” of co-occurring features: Complex objects are described in terms of the chunks they are made of:
  - Words as chunks of letters; sentences as chunks of words
  - Faces as chunks of visual features (eyes, nose, mouth)
- Memory organizes information into chunks: STM capacity as 7±2 “chunks”
- We use the resulting chunks to interpret novel input:
  - Memory shapes perception; perception shapes memory

Competitive chunking
Servan-Schreiber & Anderson 1990

Perception of the string “MTVR”
Competitive chunking
Servan-Schreiber & Anderson 1990

Perception of the string “MTVR”
Competitive chunking
Servan-Schreiber & Anderson 1990

- Perception of the string “MTVR”

- The higher level first: We are conscious only of the end result of perception (the most complex chunk)
What is the status of the lower-level units? Are they conscious in some sense or not?

Neural Network approaches

(see below...)
The Zombie and Commander Data

Commander Data meets the Zombies

- Two (caricatural) strategies to account for the existence of a cognitive unconscious:

  - Zombie theories assume total duplication of functions: The cognitive unconscious (the “zombie within”) is just like the cognitive conscious, only minus consciousness.

  - Commander Data theories assume that mental life is co-extensive with consciousness: Whenever some state is representational, it is also a conscious state.

- Both accounts are rooted in the “classical” notion that cognition consists of symbol manipulation.

- Neither account offers consciousness a clear computational function. It is a pure epiphenomenon.
Symbolic and subsymbolic models

Classical models are unable to accommodate implicit cognition

Zombie theories ...

- Zombie theories assume the existence of distinct but equally powerful conscious and unconscious systems:
  - Your zombie is just like your more familiar conscious self, only minus consciousness (see Searle)
  - Zombie theories are best instantiated by classical information-processing systems, which assume that the exact same sorts of computations apply with or without consciousness
  - Zombie theories thus ascribe no function to consciousness, which is merely epiphenomenal
Commander Data theories ...

- Star Trek’s character Data is an android whose bodily and cognitive innards are fully transparent to himself:
  - Except in rare circumstances (which systematically tend to be described as the result of some sort of dysfunction), Data is thus capable of describing in uncanny detail each and every aspect of its internal states: How much force he is applying when attempting to pry open a steel door, how many circuits are currently active in his positronic brain, etc.

- Commander Data theorists likewise assume that cognition is fully transparent, that is:
  - that whatever knowledge is expressed through behavior is also transparently available to introspection, that is, that consciousness allows access, with sufficient effort or attention, to all aspects of our inner lives except for those processes that are “merely” neural
  - Such theories ultimately paint a picture of cognition where knowledge is either conscious or simply not mental. C plays no clear functional role in Commander Data theories because it is assumed to be present whenever cognition happens.

Theories of implicit learning

- Implicit learning data and other phenomena of implicit cognition (i.e. blindsight, subliminal priming, etc.) have been radically reinterpreted over the past decade:
  - Instead of assuming unconscious and automatic learning of abstract concepts (the “abstractionist” perspective; rules), performance can be accounted for by episodic conscious knowledge of non-abstract information (i.e., fragments, instances; statistics): see Brooks, Whittlesea, Shanks
  - According to this perspective, the notion of implicit mental representation is superfluous.
Theories of implicit learning

- Both perspectives are profoundly unsatisfactory:
  - The notion of a full-blown cognitive unconscious that is just the same as the conscious system only minus consciousness (zombie theories) is as unsatisfactory as the notion that all of cognition involves consciousness (commander data theories)
  - Both perspectives are rooted in the classical metaphor of mind. Classical systems are modeled after conscious cognition:
    - Representations are propositional & binary
    - Representations are static and cannot be causally efficacious unless “accessed” by the processor
  - There is an alternative (subsymbolic) framework

Assumptions of direct mapping

- “Classical” theories of mind are transparent
- Sensitivity to some regularity entails that this regularity is represented as such in the cognitive system
  - Lawful behavior without rules
  - Abstract knowledge out of the processing of exemplars
- Double dissociations imply architectural modularity
  - False on logical grounds (Dunn & Kirsner, 1988)
Principle 1

Emergent representation:

- Sensitivity to some regularity does not necessarily imply that the regularity itself is represented within the cognitive system as an object of representation

- The power of simple associative learning mechanisms:
  - Lawful behavior without rules
  - Abstract knowledge out of the processing of exemplars

- Study 1: Hinton’s family trees problem
- Study 2: Sequence learning
- Study 3: Lee’s experiments
- Study 4: Object permanence
- Study 5: Learning orthographic regularities

Principle 2

Functional specialization:

Tasks are not process-pure. Modules are emergent — the product of learning rather than the beginning. Double dissociations do not imply architectural modularity

- Single-system accounts of double dissociations
- The need for separate memory systems motivated by computational principles

- Study 1: Deep dyslexia
- Study 2: Sequence learning
- Study 4: Catastrophic interference
Principle 3

- Gradedness:
  Many aspects of information processing involve graded and continuous dimensions rather than dichotomous relationships
  - Stage-like development out of continuous changes
  - Abstraction as a continuum
  - Consciousness as a continuum
  - Graded double dissociations

Two central issues

1. The role of consciousness:
   - Is cognition without consciousness possible? In what sense?

2. Knowledge representation:
   - How is abstract knowledge represented?

<table>
<thead>
<tr>
<th></th>
<th>Abstract rules</th>
<th>Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conscious knowledge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unconscious knowledge</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where is implicit learning?
Two central issues

1. **The role of consciousness:**
   - Is cognition without consciousness possible? In what sense?

2. **Knowledge representation:**
   - How is abstract knowledge represented?

<table>
<thead>
<tr>
<th></th>
<th>Abstract rules</th>
<th>Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conscious knowledge</td>
<td>explicit learning</td>
<td></td>
</tr>
<tr>
<td>Unconscious knowledge</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where is implicit learning?

Implicit learning is unconscious symbol manipulation
Two central issues

1. The role of consciousness:
   ♦ Is cognition without consciousness possible? In what sense?

2. Knowledge representation:
   ♦ How is abstract knowledge represented?

<table>
<thead>
<tr>
<th></th>
<th>Abstract rules</th>
<th>Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conscious knowledge</td>
<td>explicit learning</td>
<td>“Data”</td>
</tr>
<tr>
<td>Unconscious knowledge</td>
<td>Zombies</td>
<td></td>
</tr>
</tbody>
</table>

Implicit learning is conscious learning of associations

Many theories of implicit learning...

... tend to question its existence:

♦ O’Brien & Opie (1999) claim that all stable representations result in phenomenal experience and deny the possibility for causally efficacious but unconscious representations.

♦ Shanks & StJohn (1994) claim that the distinction between implicit and explicit learning is one of information-processing, not of consciousness, and that “Human learning is almost invariably accompanied by conscious awareness” (p. 394).

♦ Perruchet & Vinter (in press), while not denying implicit learning altogether, consider that unconscious influences on behavior should be exclusively ascribed to noncognitive, neural processes: Representations are always conscious.
Two central issues

1. The role of consciousness:
   - Is cognition without consciousness possible? In what sense?

2. Knowledge representation:
   - How is abstract knowledge represented?

<table>
<thead>
<tr>
<th></th>
<th>Abstract rules</th>
<th>Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conscious knowledge</td>
<td>explicit learning</td>
<td>“Data”</td>
</tr>
<tr>
<td>Unconscious knowledge</td>
<td>Zombies</td>
<td>Priming+</td>
</tr>
</tbody>
</table>

Implicit learning involves relational priming based on functional similarity

Measuring Awareness

Three challenges:

- What exactly is consciousness? It is not a single thing, but instead has many aspects: awareness of a stimulus vs. awareness of an intention vs. awareness that my behavior is influenced by some previous episode, etc.

- What are appropriate measures of consciousness? Simple dissociation logic is not sufficient

- Are my measures (of awareness and of performance) equally sensitive?
Measuring Awareness

Verbal reports

- Main basis for early dissociation findings and claims that learning is implicit
- BUT: Verbal reports fail to satisfy information and sensitivity criteria (Shanks & StJohn 1994). There is also the retrospective assessment problem.

Subjective tests

- Confidence judgments (Cheesman & Merikle, 1984; Dienes et al., 1995) indicate that performance on direct tests do not correlate very well with people’s own subjective appraisal of their performance: You can think you’re guessing, yet perform above chance on a direct test
- BUT: Expressing confidence judgments is difficult and left open to interpretation

Forced-choice, objective tests

- Performance on direct tests such as Recognition or Generation tasks typically correlate very well with choice reaction time performance
  - Recognition tasks require subjects to judge whether sequence fragments are familiar or novel
  - Generation tasks require subjects to produce sequences that resemble the training sequence
- Such associations have been taken as evidence that sequence learning is in fact explicit (Perruchet & Amorim, 1992; Shanks & Johnstone, 1998)
- BUT: No task is process-pure. No task can be simultaneously exhaustive and exclusive in measuring explicit knowledge. Objective tests can thus be contaminated by unconscious influences, and could therefore overestimate conscious knowledge.
Hinton (1986): Family trees

Two similar family trees representing the relationships between the members of an English family and the members of an Italian family.

The problem: What can we learn about the structure of the trees by answering questions such as:

- “Who is the son of James?”
- “Who is the mother of Victoria?”
The network

- The network is trained on many examples by
  - activating the corresponding input units
  - letting activation propagate through the network
  - presenting the correct target responses
  - using back-propagation to modify the connection weights

Internal representations

- The network spontaneously discovers important dimensions of the problem: Some hidden units encode nationality (units 1 and 5) or age (units 2 and 3)
- The network can generalize in a limited manner to novel examples
- Network generalization is related to the presence of a bottleneck that forces the network to redescribe the problem in a compressed manner (It is rather unusual to find individual hidden units that come to act as detectors for a particular aspect of the input)
- the network's performance is based exclusively on the functional similarity between training examples: Abstract representations emerge out of the processing of specific exemplars
The SRN model
(Elman, 1990)

TASK IS PREDICTION
On each time step:

- Present Element t over input units
- Copy hidden units activation onto context units
- Let activation propagate
- Compare response and actual successor ➔ error
- Modify the weights using back-propagation
Elman (1990)

- Training on a continuous sequence of random syllables:
  \[ \text{ba dii guuu} \]

- The network exhibits sensitivity to the predictability of successive elements:
The prediction error is high for (unpredictable) consonants and lower for (predictable) vowels

- Chunking without chunks!

Learning in the SRN

- Initially, the context is completely random and provides no information. The first thing the network does is to develop stable representations of the associations between the current element and its successors.

- Once each element is represented by a stable pattern of activation over the hidden units, the network can start predicting based on two elements: the current element, and the previous element, as represented over the context units. Hidden units can now start to represent the associations between two elements and their successors.

- Sensitivity to the constraints set by previous elements continues to develop in a similar way until the network ends up producing prediction responses that reflect the optimal conditional probability of each element given the temporal context. These responses can be interpreted as reflecting implicit preparation for the next element.
Discovering lexical categories

- 12 classes, 29 words
- Training on 1000 "sentences" of 2-3 words:
  WOMAN SMASH PLATE CAT MOVE
  MAN BREAK CAR BOY MOVE GIRL
  EAT BREAD DOG MOVE MOUSE
  MOUSE MOVE BOOK LION EAT

Table 3: Categories of lexical items used in sentence simulation

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN-HUM</td>
<td>man, woman</td>
</tr>
<tr>
<td>NOUN-ANIM</td>
<td>cat, mouse</td>
</tr>
<tr>
<td>NOUN-AGRESS</td>
<td>dragon, monster</td>
</tr>
<tr>
<td>NOUN-AGGRAV</td>
<td>glass, plate</td>
</tr>
<tr>
<td>NOUN-FOOD</td>
<td>cookie, sandwich</td>
</tr>
<tr>
<td>VERB-TRANCE</td>
<td>think, sleep</td>
</tr>
<tr>
<td>VERB-TRANS</td>
<td>see, chase</td>
</tr>
<tr>
<td>VERB-AGG</td>
<td>move, break</td>
</tr>
<tr>
<td>VERB-ONE</td>
<td>small, see</td>
</tr>
<tr>
<td>VERB-DESTROY</td>
<td>break, smash</td>
</tr>
<tr>
<td>VERB-ERT</td>
<td>eat</td>
</tr>
</tbody>
</table>

Internal representations

- The network acquires internal representations that reflect the hierarchical structure of the domain
Learning Finite-state grammars
Cleeremans et al. (1990)

Difficult prediction problem:
- Each letter appears twice in the grammar
- Different occurrences of the same letter can be followed by different successors

Training and testing

Training:
- on strings generated randomly from the grammar
- 60,000 trials involving strings ranging from 3 to 30 letters

Testing:
- 20,000 random grammatical strings
  - The network accepts all of them (activation of successor > 0.3)
- 130,000 random strings
  - 0.2% grammatical, 99.8 non-grammatical
  - The network correctly classifies all the strings
- Several very long strings (length > 150 letters)
  - The network correctly predicts the grammatical successors
Internal representations

- Cluster analysis on a small random set of grammatical strings.
- The network discovers the nodes of the grammar (This is the optimal thing to do!)
- Emergence of abstract knowledge driven by functional similarity (as in the Hinton network)
- Important caveats:
  - Some of the structure is already present before training
  - Node discovery only occurs when the number of hidden units is very small

Conclusions

- The network can learn the grammar PERFECTLY
- It can become a perfect finite-state recognizer, able to accept or reject any presented sequence
- Crucial implication: The network can learn to generalize to an infinite set of strings based on (necessarily finite) exposure to a set of examples
- The network strikes a balance between push-down automata and finite-state machines

SRNs are GRADED STATE MACHINES
Sequence Learning

Training

Generation

Inclusion:
“Produce the next element”
Implicit and explicit cooperate

Exclusion:
“Produce anything but the next element”
Implicit and explicit influences act in opposition

Typical Results

Implicit Learning:
A change in performance that is not accompanied by a corresponding change in the ability to describe the acquired knowledge
Sequence Learning

- A sequence is generated based on an FSG
- Random stimuli are randomly inserted in the sequence
- If people learn something about the grammar, they should respond faster to predictable G stimuli than to unpredictable NG stimuli

Modeling sequence learning
Cleeremans & McClelland (1991)

R-SQUARED OVER ALL NODES : .88
Abstraction and consciousness

- The previous examples all indicate that abstract knowledge can emerge out of sensitivity to the functional similarity between training examples.
- Nevertheless, the resulting abstract knowledge remains unavailable as knowledge to the networks: They have no meta-knowledge.
- Human abstract knowledge is often characterized by the simultaneous availability of meta-knowledge.
- It remains a formidable challenge to figure out how relevant meta-knowledge can emerge out of statistical learning mechanisms (possible answer: Use activation-based processing).

Clark and Karmiloff-Smith (1993)

- Four strategies though which to understand the relationships between subsymbolic and symbolic cognition:
  1. Give up connectionism entirely and revert to a thoroughly classical approach (despair).
  2. Augment connectionist-style networks with the symbol structures of natural language (a representational leap).
  3. Combine elements of connectionism and classicism in a single system (hybridization).
  4. Use thoroughly connectionist resources in increasingly sophisticated ways (more of the same).
Lee’s experiment

Lee (1997) introduced a challenging sequence learning situation:

- Participants are trained on essentially random material that follows a single highly abstract rule:

  \[ \text{Each stimulus appears once and only once in any 6-elements sequence} \]

  1 2 3 4 5 6 5 6 1 2 3 4 2 1 4 3 5 6 ...

- Beginnings and endings of each sequence are not marked in any way

- Within each sequence, the predictability of each element increases with its serial position: The final element is always perfectly predictable based on the previous elements, and the first element is always random
Lee's results

Lee showed that participants could learn this rule quite well, producing faster responses to serial position 6 stimuli than to serial position 1 stimuli.

Lee’s material presents an interesting challenge to theories of sequence learning:

- Instance-based accounts are in trouble: There is no repeating pattern to latch onto.
- Fragment-based accounts are in trouble: It is hard to see which fragments should be encoded, as all possible fragments occur and can end in all serial positions.
- Frequency-based accounts are in trouble: All elements have the same frequency.
- Pattern-based accounts are in trouble: Most target movements tend to occur and, they fail to instantiate informative regularities.

Lee's conclusions

Based on these findings, Lee concluded that:

- “both parsing and short-term memory mechanisms are necessary”

... and that association-based models such as the SRN could not learn the material.

Hmm!
A conceptual replication

- Serial Reaction Time Task:
  - 6 choices
  - Training on the entire stimulus set of 720 (6!) sequences presented over 24 blocks of 180 trials each, for a total of 4320 trials
  - Neutral implicit learning instructions: “Respond as fast and as accurately as you can”; “this experiment is about motor performance”

Predictions

- Reaction time performance is related to the statistical structure of the material: Predictable stimuli should elicit faster reaction times than unpredictable stimuli
Reaction Time Performance

- Participants exhibit clear sensitivity to the statistical structure of the material, yet …
- … claim that the material is completely random

Participants do not learn much

- We failed to find strong evidence of learning in this situation: The serial position effect emerges very early in training and remains relatively stable over time
- The absence of learning makes the serial position effect even more intriguing:
  - Where does it come from?
  - How can participants be sensitive to the structure of the material even early in training?
Where does the serial position effect come from?

- The “parsing” account:
  - According to Lee (1997), participants respond faster to serial position 6 stimuli than to serial position 1 stimuli because the former are more predictable than the latter.
  - This account, while correct, makes it mandatory to assume that participants are somehow able to implicitly parse or segment the material into the correct chunks, that is, to know what the current serial position is.
  - This appears to be highly implausible in the current situation.

The serial position effect is emergent

- The “repetition distance” or “lag” account:
  - The serial position effect is emergent and does not require that participants somehow parse the stimulus material.
  - Instead, they simply tend to respond faster to stimuli that are associated with a long lag, that is, stimuli that have not occurred recently.
  - By construction, the stimulus material is such that longer lags are associated with late serial positions:
    - Any stimulus that occurs on serial position 6 of an experimenter-generated sequence has by definition not occurred within the last 5 trials.
    - In contrast, a stimulus that occurs on serial position 1 could have occurred as recently as 2 trials ago.
The lag effect

The serial position effect observed by Lee (1997) and replicated in this experiment is in fact entirely reducible to a "repetition distance" or "lag" effect such that:

- on each trial, the probability of any stimulus increases linearly with the lag that separates the current stimulus from the same stimulus’ previous occurrence
- late serial positions in the experimenter-generated sequences are, by construction, associated with distributions of increasingly longer lags

The lag effect in the human data

Each curve (corresponding to stimuli with a specific lag) tends to be flat across serial positions

Reaction times do not appear to be influenced by serial position, but rather by the lag that separates the current stimulus from its previous occurrence
A simple model of the lag effect

To find how whether sensitivity to repetition distance is sufficient to account for performance, we constructed a simple model that implements this sensitivity directly:

The SIMPLE CONDENSATOR MODEL

The simple condensator model:

- Six units, each corresponding to a given response
- Each unit is associated with an activation level ranging from 0 to 1
- When a stimulus is presented, the corresponding response unit fires and its activation is divided equally among the other 5 units
- During processing, the activation of each unit increases with time up until the point when the corresponding stimulus is presented: it's clear that the dynamics of the SCM depend directly on repetition distance: the most active unit always corresponds to the stimulus associated with the longest lag
The model’s responses correspond almost perfectly with the actual distribution of lags.

SCM responses decrease linearly from the first to the tenth lag.

What is the origin of sensitivity to repetition distance?

- The SCM model, like others assumes that sensitivity to repetition distance is “hardwired”
- An alternative hypothesis is that sensitivity to repetition distance is a result of the interaction of prediction-driven mechanisms with environments in which repetition distance has predictive value.
What is the origin of sensitivity to repetition distance?

- In the real world, events do not occur randomly, but repeat at different predictable intervals: Repetition distance has predictive value
  - The probability of your having a meal is close to zero if you’ve just had one, and increases monotonically with the time since you had your last meal
  - The same situation holds for myriads of other real-world events
- Sensitivity to repetition distance is a learned bias with which participants enter the experiment
- This bias emerges out of more elementary prediction-based associative learning mechanisms

What is the origin of sensitivity to repetition distance?

- Sensitivity to repetition distance contrasts with assumptions in ACT-R and with data (Schooler et al.) showing how memory fulfills need functions
- Other data suggest that sensitivity to repetition distance, or negative recency, is important:
  - The gambler's fallacy
  - Inhibition of return
  - Negative priming
- Can sensitivity to repetition distance be learned?
The Simple Recurrent Network

Can the SRN model master this material?
- The SRN fails to master the material in a single epoch
- However, it succeeds if previously trained on a noisy version of the material for 6 epochs
- This pre-training is assumed to reflect previous experience with the real-world

SRN Simulation results

Like human participants, the networks exhibit a linear serial position effect
SRN Simulation results

Over training, the network learns to implement a Simple Condensator Model: It activates most strongly those units that correspond to responses that have not been produced recently.

The lag effect in the SRN

Like human participants, the network appears to be insensitive to serial position as such, and instead produces responses that depend essentially on the lag.
The SRN account

- The serial position effect described by Lee (1997) emerges out of a more elementary sensitivity to the lag that separates two occurrences of the same stimulus.

- The lag effect itself emerges out the network’s prediction-based sensitivity to the statistical structure of the material:
  - The fragment ‘ABC’, by construction:
    - can never be followed by ‘C’
    - is more often associated with ‘A’ as a successor than with ‘B’ regardless of the serial position at which ‘ABC’ ends, that is, the fragment ‘ABCA’ occurs more frequently (e.g., A-BCA & AB-CA & ABC-A can occur) than the fragment ‘ABCB’ (e.g., only AB-CB & ABC-B can occur).

Conclusions

- Abstract knowledge is not involved in this situation:
  - The serial position effect, rather than reflecting rule-like knowledge, in fact emerges out of a simpler “repetition distance” effect.
  - Sensitivity to repetition distance in turn emerges naturally out of elementary prediction-based learning systems interacting with environments in which repetition distance has predictive value.
  - Learned real-world biases to preferentially prepare for responses that have not been produced recently play a significant role in sequence learning.
Cognitive Development

- Two main paradigms:
  - Habituation-dishabituation paradigm:
    Infants are exposed to a set of stimuli until bored. New stimuli are then presented, and attention to them measured.

  - Violation-of-expectation paradigm:
    Infants are familiarized with an event, and then shown novel events that are either possible or not. The impossible event is physically similar to the familiar event, but conceptually different from it. The reverse holds for the possible event. Longer looking at the impossible event is taken as an indication that infants have the relevant concept.
Baillargeon Experiments

Object permanence involves the understanding that objects continue to exist independently of our percepts of them.

- Early characterization (Piaget): Object permanence is acquired late in development (8 months….) based on reaching behavior.

- More recent characterization (Baillargeon, Spelke): Object permanence is acquired early (3.5 months) based on looking time.

- The discrepancy is traditionally accounted for by the following two assumptions:
  - Object permanence is an all-or-none concept (it is propositional and innate).
  - Different measures (e.g., preference, reaching) have different sensitivity to that knowledge: The reaching deficit stems from inadequate means-ends analysis. Infants have the concept of object permanence (as made clear by the looking time data) but fail to be able to organize an appropriate action sequence to reach for an occluded object.
Graded knowledge

- Knowledge of concepts such as object permanence is not all-or-none, but graded: it strengthens with time, and is embedded in specific processing modules. The more modules it is embedded in, the wider the range of behaviors that can make use of the knowledge.

- Looking time is based on perceived mismatch between constantly operation prediction-based expectations and actual events.

Graded representation theory

Looking time and reaching are both supported by the same gradual knowledge, and are associated with different thresholds.
Testing the theories

- Use tasks that require the same means-ends abilities, but different object permanence knowledge:
  - Bower & Wishart (1972) contrasted tasks requiring reaching for objects located inside a box with either a transparent or an opaque cover. Three months-old infants successfully retrieve objects in the transparent boxes but not objects in opaque boxes.
  - However, it could still be possible that infants have relevant object knowledge, but fail to know how to retrieve objects. Indeed, when the cover is transparent, the object is visible and infants spontaneously reach for it. They could thus accidentally lift the cover or grasp it, even though they have no real knowledge of how to lift the cover. The two situations would thus not, in fact, require the same means-ends analysis.
  - Use tasks in which no advantage is given through direct reaching in the transparent condition.

Experiment 2
Munakata et al. (1997)

- Infant are trained to push a button so as to retrieve an object located behind a screen.
- Four conditions: Object (present or not) X screen (opaque or transparent).
- Dependent measure: # of button presses.
More retrievals in the transparent condition
• The difference cannot be attributed to a deficit in means-end analysis because means-ends analysis demands are the same in both conditions
• The difference therefore stems from infants’ weaker representations of objects when they are occluded

Results

Conclusions

• The findings call into question the notion that object permanence is all-or-none and innate or present very early on:
  ♦ Infants’ representations of occluded objects can vary in strength and in accessibility to output systems
  ♦ Infants can succeed on looking-time but not reaching independent of deficits in the execution of reaching
Simple recurrent networks are trained to predict the future position of objects. The recurrent connections allow the networks to use information about the past in order to predict the future. Successful predictions when objects can be occluded require the development of internal representations that continue to exist when the object is no longer visible.

Simulation results

- Sensitivity = the network’s expectation for the ball’s reappearance when there was a ball behind the occluder minus network’s expectation for the ball’s reappearance when there was no ball behind the occluder.
- The network becomes increasingly capable of predicting the reappearance of an object after occlusion, and after longer and longer intervals of occlusion.
- It generalizes to never-seen instances (using distributed inputs).
Internal representations

- Use subtraction (just like brain imaging) to figure out which internal units code for what
- Units 1, 8, 10, 11, and 15 code for the ball most strongly
- After 100 epochs of training, the network is not really able to maintain the representation of the ball when it is occluded

More training...

- Internal representations of the occluded ball become stronger and remain stronger for longer periods with training
The reaching system is delayed for occluded objects, but can respond appropriately to visible objects early in training (not shown)

Confirms the notion that strength of representation is the crucial factor

Conclusions

Munakata’s “adaptive process” theory assumes that:

- The connections that underly both the development and the use of representations needed to perform object-permanence tasks strengthen with experience
- Reaching requires stronger representations
- Simulations demonstrate how a “graded” account of the development of object permanence concepts can capture the data

This stands in contrast with traditional theories of object permanence, which assume an early, all-or-none sensitivity

The results raise interesting issues about the unity of consciousness
Cognitive control

- The task is to sort cards according to either their color or their shape.
- When asked to classify according to shape, children perseverate in using color even though they are quite capable of indicating how cards should be classified according to shape!

>> Study 5
Implicit Learning out of the lab
Implicit learning in the real world

- Most studies of implicit learning are limited to laboratory settings, yet all authors claim that implicit learning is ubiquitous in the real world (for instance, IL is often assumed to be essential in language learning)

- The goal of this study was to explore how sensitivity to orthographic regularities develops over very long (i.e., years) periods of time so as...

- … to enable us to contrast rule-based approaches to language learning with statistical approaches (Marcus/Pinker vs. Christiansen, Seidenberg, Elman, McClelland, etc.) in a way that is both ecologically valid and impossible to implement in the lab

Transfer situations

- Implicit learning transfer situations all involve training subjects on material instantiated in a given vocabulary of surface features and testing them on material instantiated in a different vocabulary
  - Different letter sets in Artificial Grammar learning
  - Different modalities

- Performance on the novel material is often better than chance, but systematically worse than on the training material: this decrease is called transfer decrement
Theories of transfer

- Two “traditional” approaches:
  - Abstractionist, or rule-based approaches (assume surface independence)
  - Distributional (statistical) approaches (depend on surface similarity)

- But there are two orthogonal dimensions
  - Independence from surface features
  - Rule-based character of knowledge representations

- Hence:
  - Rules can involve surface features
  - Memory and statistical processes can operate on abstract features: Crucial notion of “functional or abstract similarity” in connectionist models

Contrasted predictions

- Abstractionist approaches ultimately (i.e., after sufficient practice) predict perfect transfer to material instantiated with different surface features

- Distributional approaches predict the persistence of transfer decrement over practice because the similarity between training and test materials does not change with practice
Learning of orthographic regularities

- These predictions can not be tested in the lab because it can always be argued that necessarily limited practice fails to offer sufficient opportunities for the emergence of pure abstract knowledge.

- To address this issue, we examined the development of children's knowledge of orthographic regularities over five years of exposure: Language learning as a giant artificial grammar learning experiment!


Goal

- Some orthographic regularities in French can be expressed as rules:
  - Vowels can never be doubled
  - Some consonants can be doubled and other cannot
  - Double consonants can never appear at the beginning or end of a word

- Use of never-doubled consonants allow us to assess whether children's sensitivity to the regularities involves abstraction

- The task was non-word preference judgment:
  - Do children prefer non-words containing a doublet of never-doubled consonants over non-words containing a double vowel?
  - Does knowledge about the legal positions of double consonants extend to never-doubled consonants?
  - How does performance change over YEARS of exposure?
Design

- Material
  - 12 pairs of 6-letter non-words comparing frequently doubled consonants and never-doubled consonants (CCf/CCn pairs):
    ullate - ujjate
  - 24 pairs of 6-letter non-words comparing double vowels and either frequently or never-doubled consonants (VV/CCf and VV/CCn pairs):
    buumer - bummer          ahhire - ahiire
  - 24 pairs of 6-letter non-words comparing the position at which either frequently or never-doubled consonants appear (Position CCf and Position CCn pairs):
    nnulor - nullor       rixohh - rixxoh
- Participants: Five groups of children aged 6-11

Results

- VV-CCf: Children prefer items containing frequently doubled consonants over items containing double vowels
- CCf vs. CCn: Children prefer items containing frequently doubled consonants over items containing never doubled consonants
- VV-CCn: Children prefer items containing never doubled consonants over items containing double vowels
  - Indicative of abstraction
- Transfer decrement remains constant across ages
  - Inconsistent with the development of rule-based knowledge
Simulations

- The observed parallelism in performance on familiar and novel material suggests that a single system is involved.

- An SRN using local representations on its input and output pools was trained to predict the next letter of each of the 1,000 most frequent words in French.

- Words were selected for presentation at random according to their frequency of occurrence in the language.

- Ten networks were trained on approximately one million words over 10,000 epochs, and tested at various intervals during training.

Simulation results

- VVs vs. Cs: The network initially predicts that vowels are more likely to be doubled than frequently doubled consonants because vowels are more frequent in general. This changes with training.

- The network becomes sensitive to the vowel/consonant distinction: With training, the network ends up preferring never doubled consonants over never doubled vowels (but the preference is less marked than for frequently doubled consonants).

\[
\text{Cf/(Cf+V)} = \text{network's average prediction that a frequently doubled consonant will be doubled divided by its overall tendency to predict doubling.}
\]
Conclusions

- The persistence of transfer decrement is indicative that transfer is based on functional similarity rather than on rule-based knowledge.
- Implicit learning actually occurs in the real world!
Overall conclusions

- Abstraction is graded, at least over the larger portion of a continuum extending from pure instance-based processing to fully abstraction-based processing.

- Emergence: Sensitivity to some regularity does not necessarily imply that the regularity itself is represented in the cognitive system as an object of representation.

- Hybridization: There is a real challenge involved in figuring out how full-fledged symbolic representations emerge out of subsymbolic processing and representational systems (distinction between weight-based and activation-based processing?)