Computational models of goal-directed behavior

Marc Toussaint

Machine Learning & Robotics Lab – FU Berlin

marc.toussaint@fu-berlin.de

Berlin, July 17th, 2010

Wolfgang Köhler (1917)

Intelligenzprüfungen am Menschenaffen

The Mentality of Apes

[movie]

Wolfgang Köhler (1917)

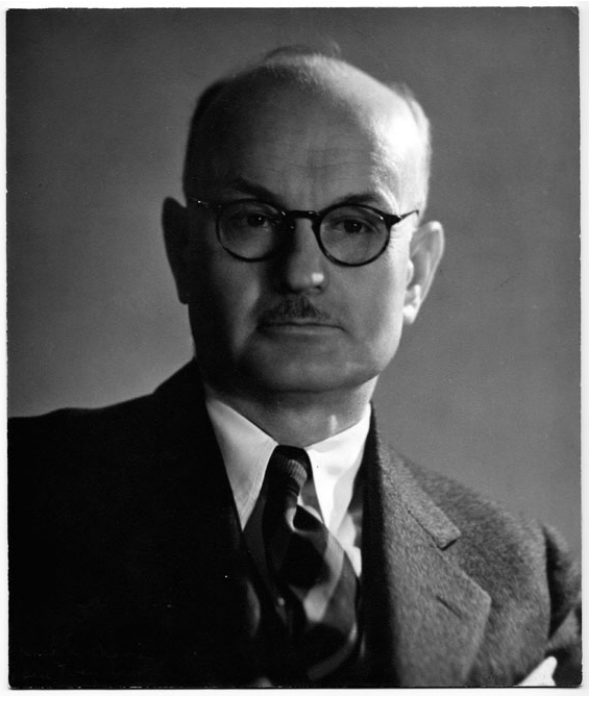
Intelligenzprüfungen am Menschenaffen

The Mentality of Apes

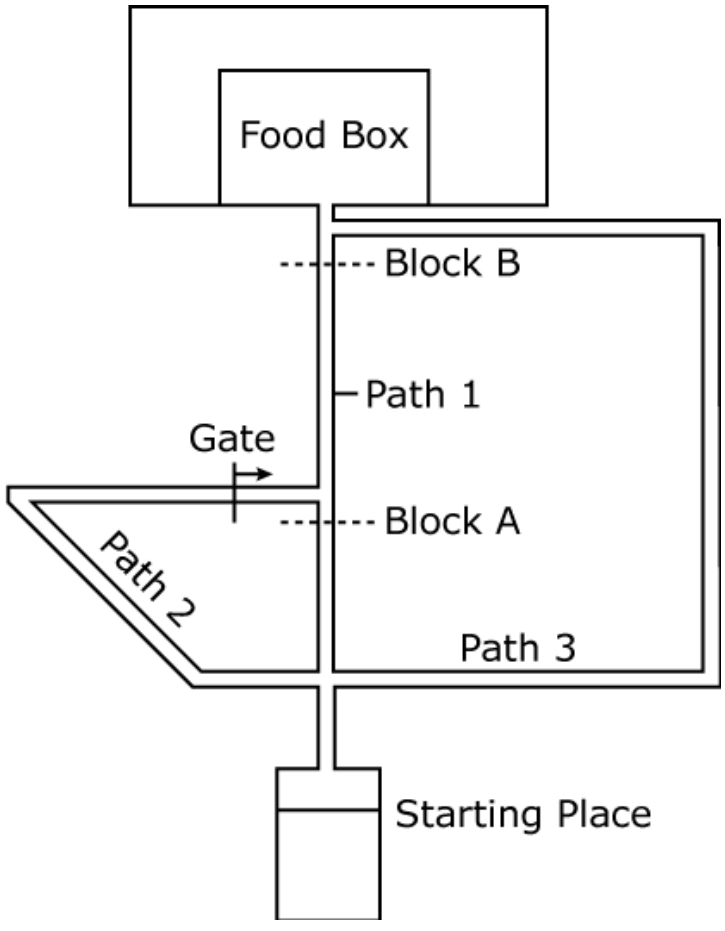
[movie]

Goal-directed Decision Making

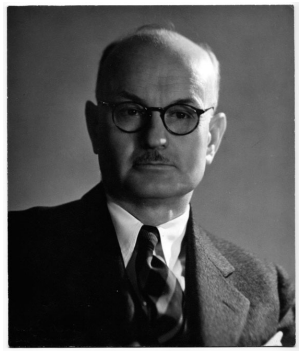
What are computational principles for such behavior?



Edward Tolman (1886 - 1959)



Model-based versus Model-free

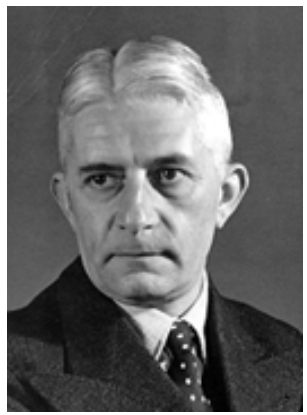


Edward Tolman (1886 - 1959)

Clark Hull (1884 - 1952)

Principles of Behavior (1943)

learn stimulus-response mappings based on reinforcement



Wolfgang Köhler (1887–1967)

learn facts about the world that they could subsequently use in a flexible manner, rather than simply learning automatic responses

experience

$\{s_t, a_t, s_{t+1}, r_t\}$

knowledge

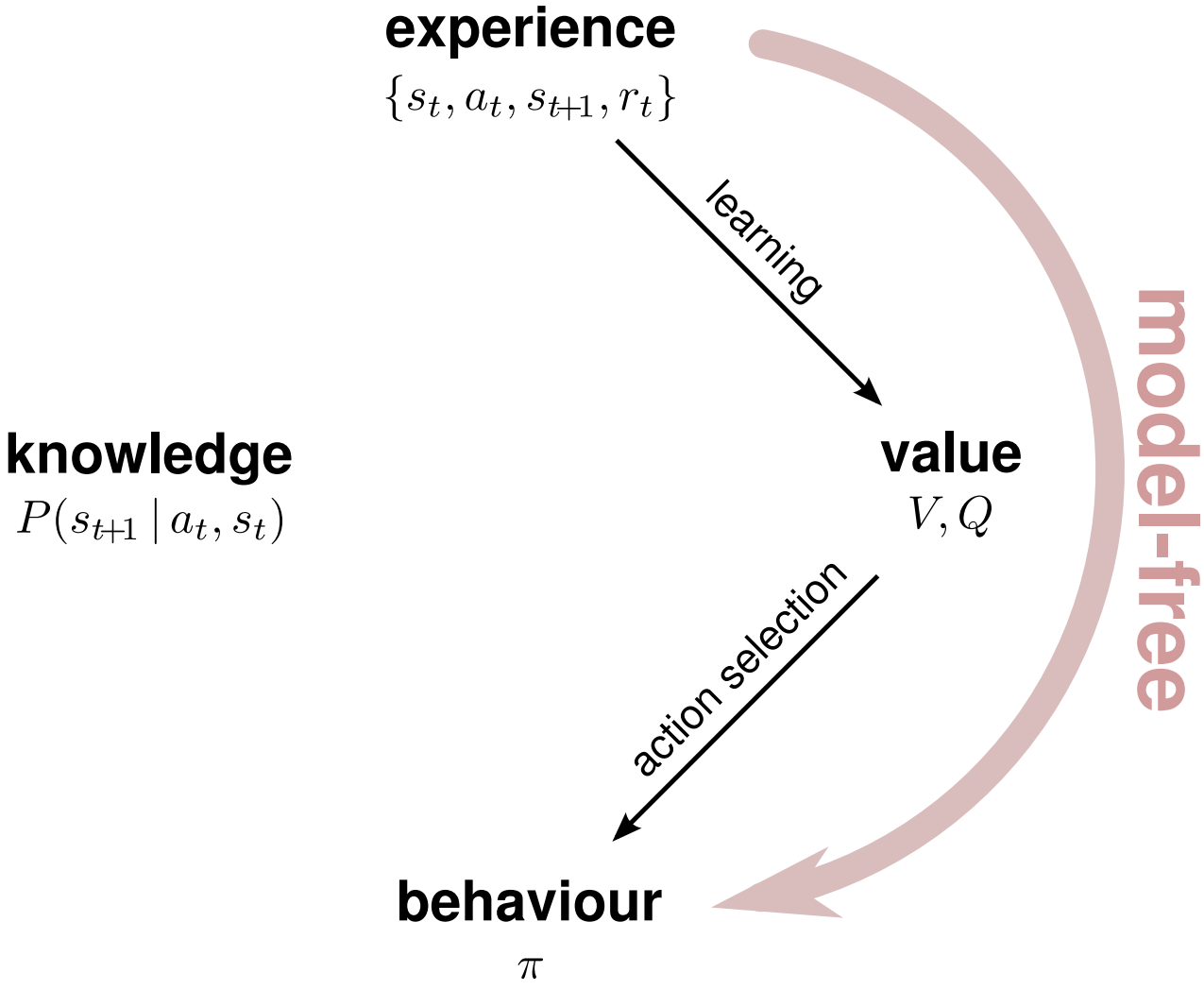
$P(s_{t+1} \mid a_t, s_t)$

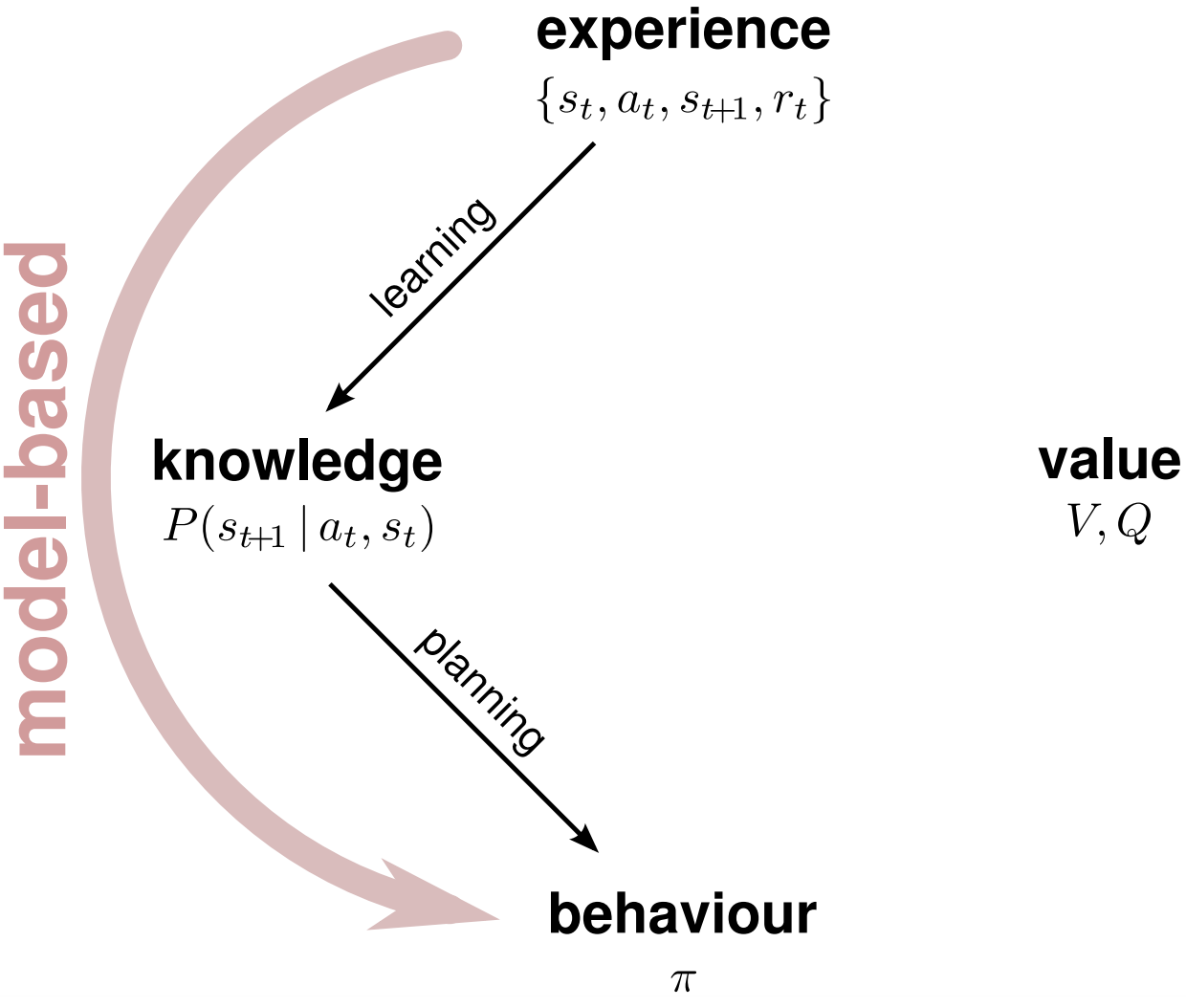
value

V, Q

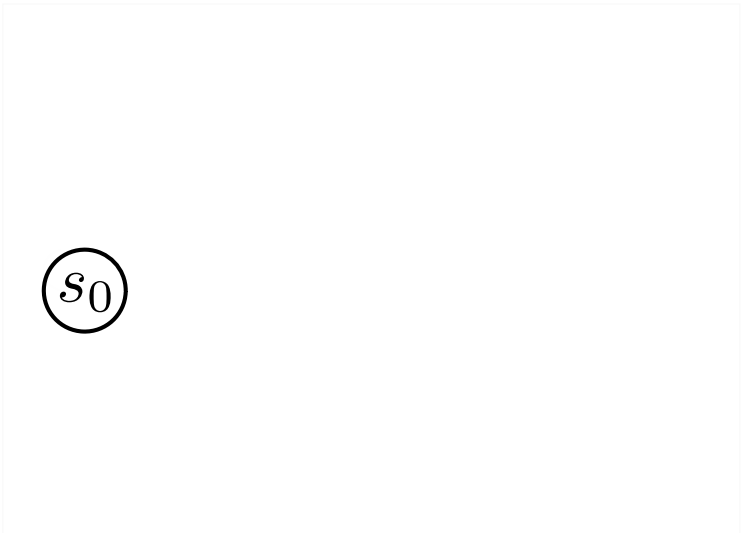
behaviour

π


experience
$\{s_t, a_t, s_{t+1}, r_t\}$
learning
knowledge
$P(s_{t+1} \mid a_t, s_t)$
value
V, Q
action selection
behaviour
π
model-free

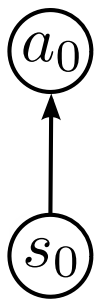

experience
$\{s_t, a_t, s_{t+1}, r_t\}$
learning
knowledge
$P(s_{t+1} \mid a_t, s_t)$
value
V, Q
planning
behaviour
π
model-based

Reinforcement Learning Basics



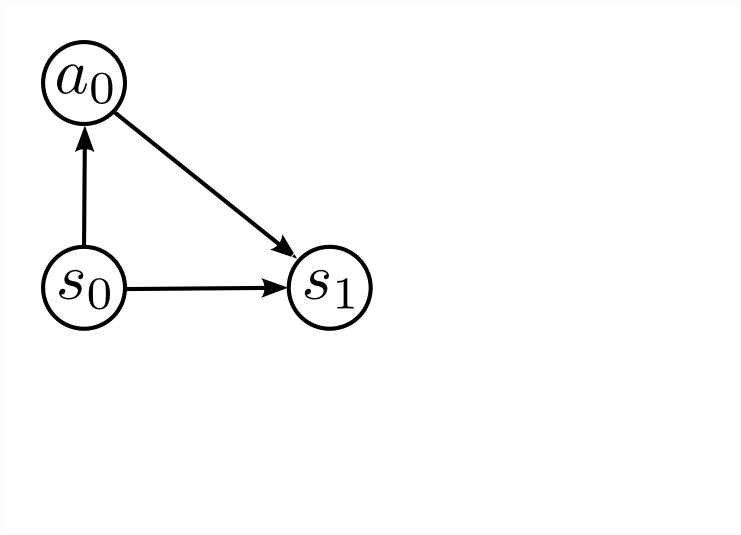
start state s_0

Reinforcement Learning Basics



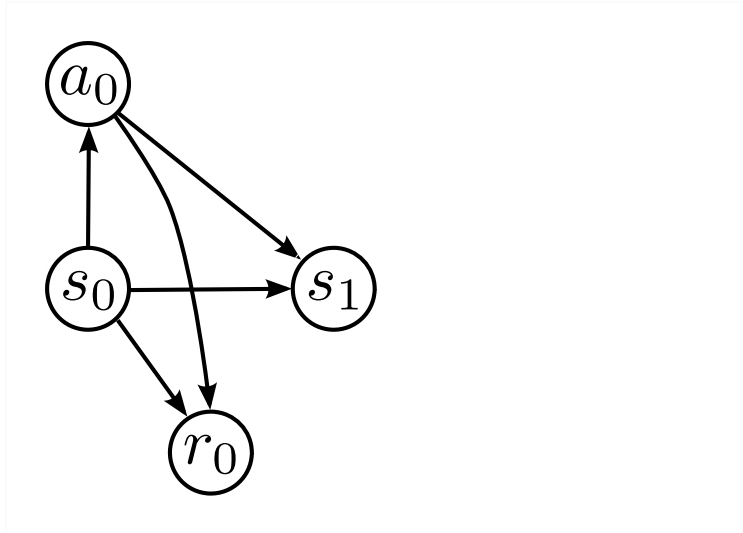
agent chooses $a_0 = \pi(s_0), \quad \pi =$policy

Reinforcement Learning Basics



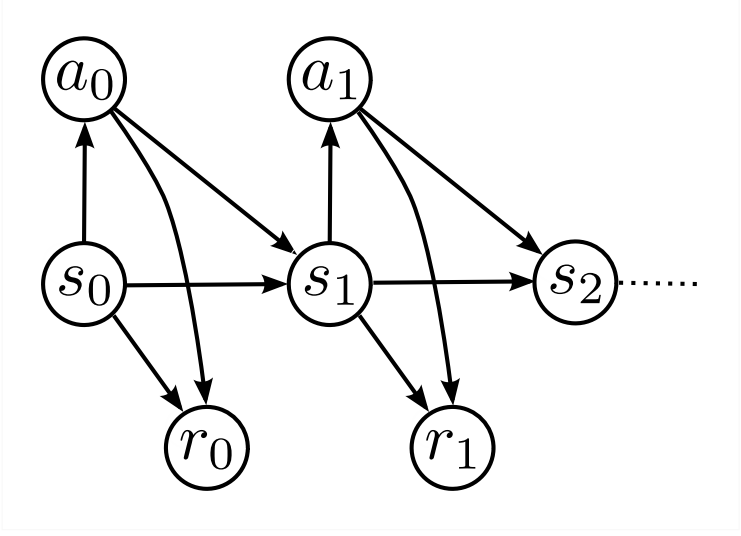
world makes a state transition $P(s_1 \mid a_0, s_0)$

Reinforcement Learning Basics



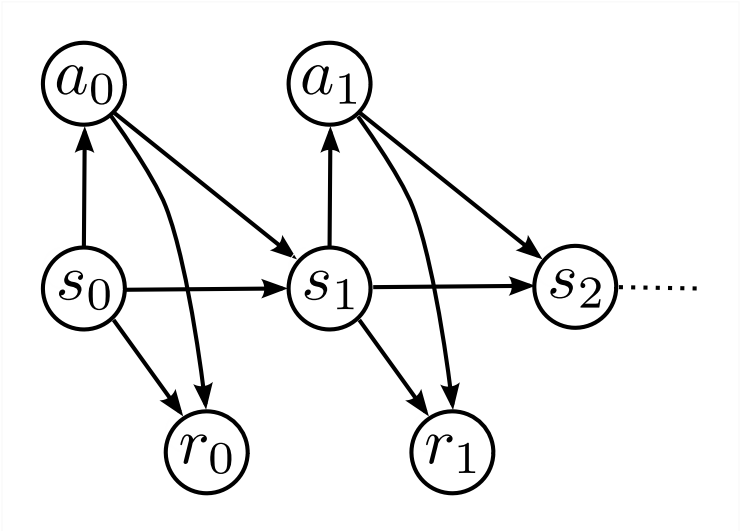
world gives a reward $P(r_1 \mid a_0, s_0)$

Reinforcement Learning Basics



iteration. . .

Reinforcement Learning Basics



- Markov Decision Process:
  – world's transition probabilities $P(s_{t+1} \mid a_t, s_t)$
  – world's reward probabilities $P(r_t \mid a_t, s_t)$
  – agent's behavior $\pi(a_t \mid s_t)$ (or deterministic: $a_t = \pi(s_t)$)

Reinforcement Learning Basics

- The **value** (*expected* discounted return) of policy π when started in state s:

$$V^{\pi}(s) = E_{\pi}\{r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots \mid s_0 = s\} \tag{1}$$

discounting factor $\gamma \in [0, 1]$

Reinforcement Learning Basics

- The **value** (*expected* discounted return) of policy π when started in state s:

$$V^{\pi}(s) = E_{\pi}\{r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots \mid s_0 = s\} \tag{1}$$

discounting factor $\gamma \in [0, 1]$

- definition of **optimality**:
behavior π^* *is optimal iff*

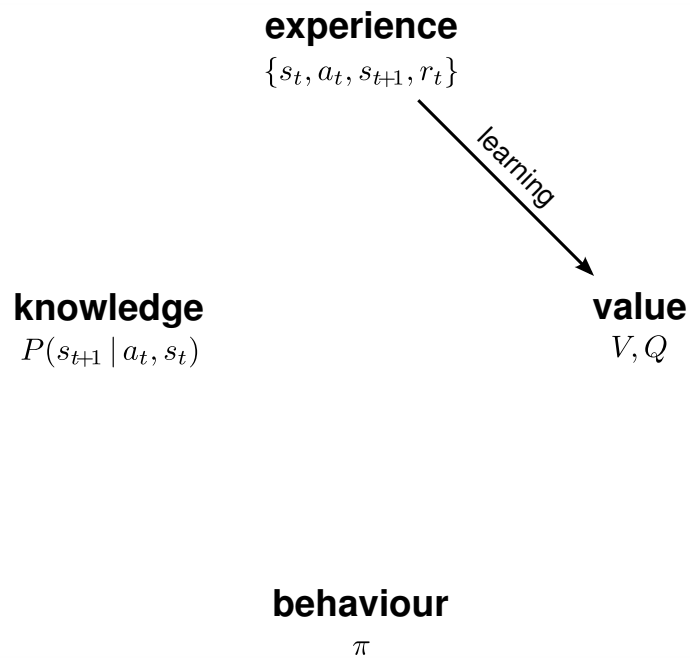
$$\forall_s : \; V^{\pi^*}(s) = V^*(s) \qquad \text{where} \;\; V^*(s) = \max_{\pi} V^{\pi}(s)$$

(simultaneously maximising the value in all states)

(In MDPs there always exists (at least one) optimal deterministic policy optimal policy.)

The prototype of model-free Reinforcement Learning:

Temporal difference learning



Temporal difference (TD) learning

- *Learnig values without a model*

Temporal difference (TD) learning

- *Learnig values without a model*

$$
\begin{aligned}
V^{\pi}(s) &= \mathrm{E}\{r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots \mid s_0 = s; \pi\} \\
&= \mathrm{E}\{r_0 \mid s_0 = s; \pi\} + \gamma \mathrm{E}\{r_1 + \gamma r_2 + \cdots \mid s_0 = s; \pi\} \\
&= \langle r_0 \rangle_{\pi, s_0} + \gamma \textstyle\sum_{s'} P(s' \mid \pi(s), s)\ \mathrm{E}\{r_1 + \gamma r_2 + \cdots \mid s_1 = s'; \pi\} \\
&= \langle r_0 \rangle_{\pi, s_0} + \gamma \textstyle\sum_{s'} P(s' \mid \pi(s), s)\ V^{\pi}(s') \qquad (2)
\end{aligned}
$$

Temporal difference (TD) learning

- *Learnig values without a model*

$$
\begin{aligned}
V^\pi(s) &= \mathrm{E}\{r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots \mid s_0 = s; \pi\} \\
&= \mathrm{E}\{r_0 \mid s_0 = s; \pi\} + \gamma \mathrm{E}\{r_1 + \gamma r_2 + \cdots \mid s_0 = s; \pi\} \\
&= \langle r_0 \rangle_{\pi, s_0} + \gamma \textstyle\sum_{s'} P(s' \mid \pi(s), s)\ \mathrm{E}\{r_1 + \gamma r_2 + \cdots \mid s_1 = s'; \pi\} \\
&= \langle r_0 \rangle_{\pi, s_0} + \gamma \textstyle\sum_{s'} P(s' \mid \pi(s), s)\ V^\pi(s') \qquad (2)
\end{aligned}
$$

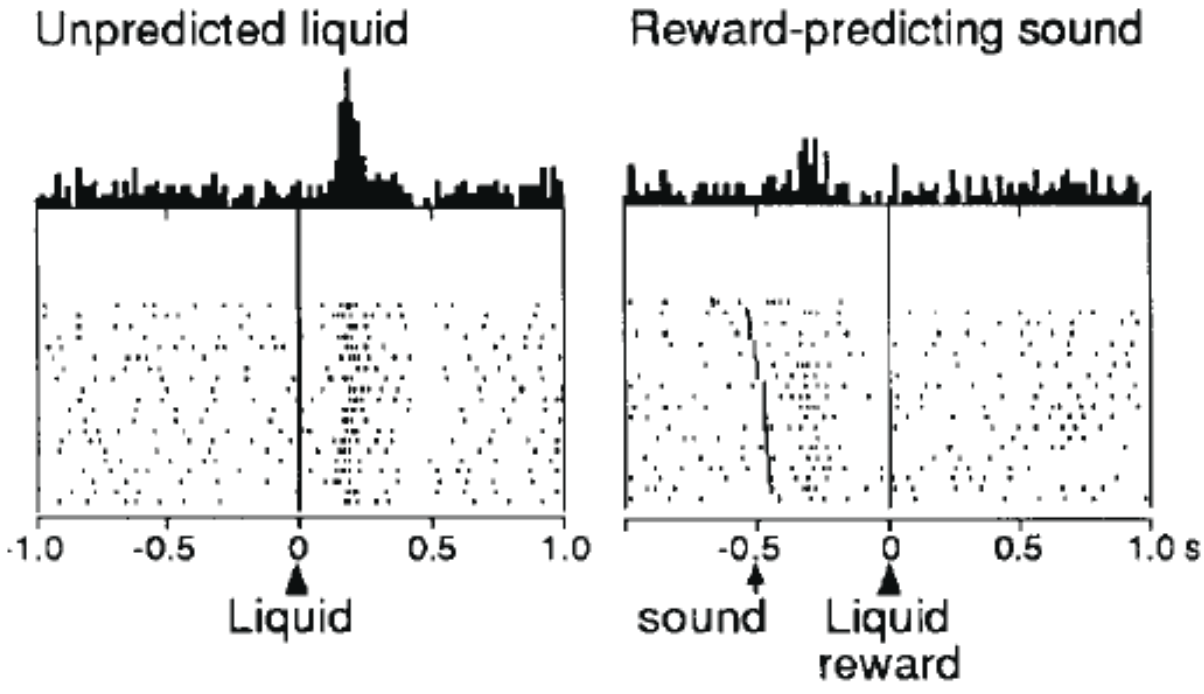
- **TD learning:** given a new experience (s_t, a_t, s_{t+1}, r_t)

$$
\begin{aligned}
V_{\text{new}}(s_t) &= (1 - \alpha)\ V_{\text{old}}(s_t) + \alpha\ [r_t + \gamma V_{\text{old}}(s_{t+1})] \\
&= V_{\text{old}}(s_t) + \alpha [r_t + \gamma V_{\text{old}}(s_{t+1}) - V_{\text{old}}(s_t)]\ .
\end{aligned}
$$

  (stochastic variant of Dynamic Programming, convergence with probability 1)

- **reinforcement:**
  - more reward than expected $(r_t > \gamma V_{\text{old}}(s_{t+1}) - V_{\text{old}}(s_t))$
    $\rightarrow$ increase $V(s_t)$
  - less reward than expected $(r_t < \gamma V_{\text{old}}(s_{t+1}) - V_{\text{old}}(s_t))$
    $\rightarrow$ decrease $V(s_t)$

Dopamine



Montague, Dayan & Sejnowski: *A Framework for Mesencephalic Dopamine Systems based on Predictive Hebbian Learning.* Journal of Neuroscience, 16:1936-1947, 1996.

so what does that mean?

so what does that mean?

– we derived an algorithm from a general framework
– this algorithm involves a specific variable (reward residual)
– we find a neural correlate of exactly this variable

great!

devil's advocate:

– does not proof that TD learning is going on
  only that an expected reward is compared with a experienced reward

– does not discriminate between model-based and model-free
  (both can induce an expected reward)

For completeness . . . Policy Iteration

– learning values is half the story

– adapting the behavior is the other half

For completeness . . . Policy Iteration

– learning values is half the story

– adapting the behavior is the other half

- Policy Iteration:

  loop

  policy evaluation using

  $V^\pi(s) \leftarrow \langle r_0 \rangle_{\pi,s_0} + \gamma \sum_{s'} P(s' \mid \pi(s), s)\ V^\pi(s')$ (or TD learning)

  policy improvement using

  $\pi(s) \leftarrow \mathrm{argmax}_a \langle r_0 \rangle_{\pi,s_0} + \gamma \sum_{s'} P(s' \mid \pi(s), s)\ V^\pi(s')$

  end loop

Limitations of the model-free view

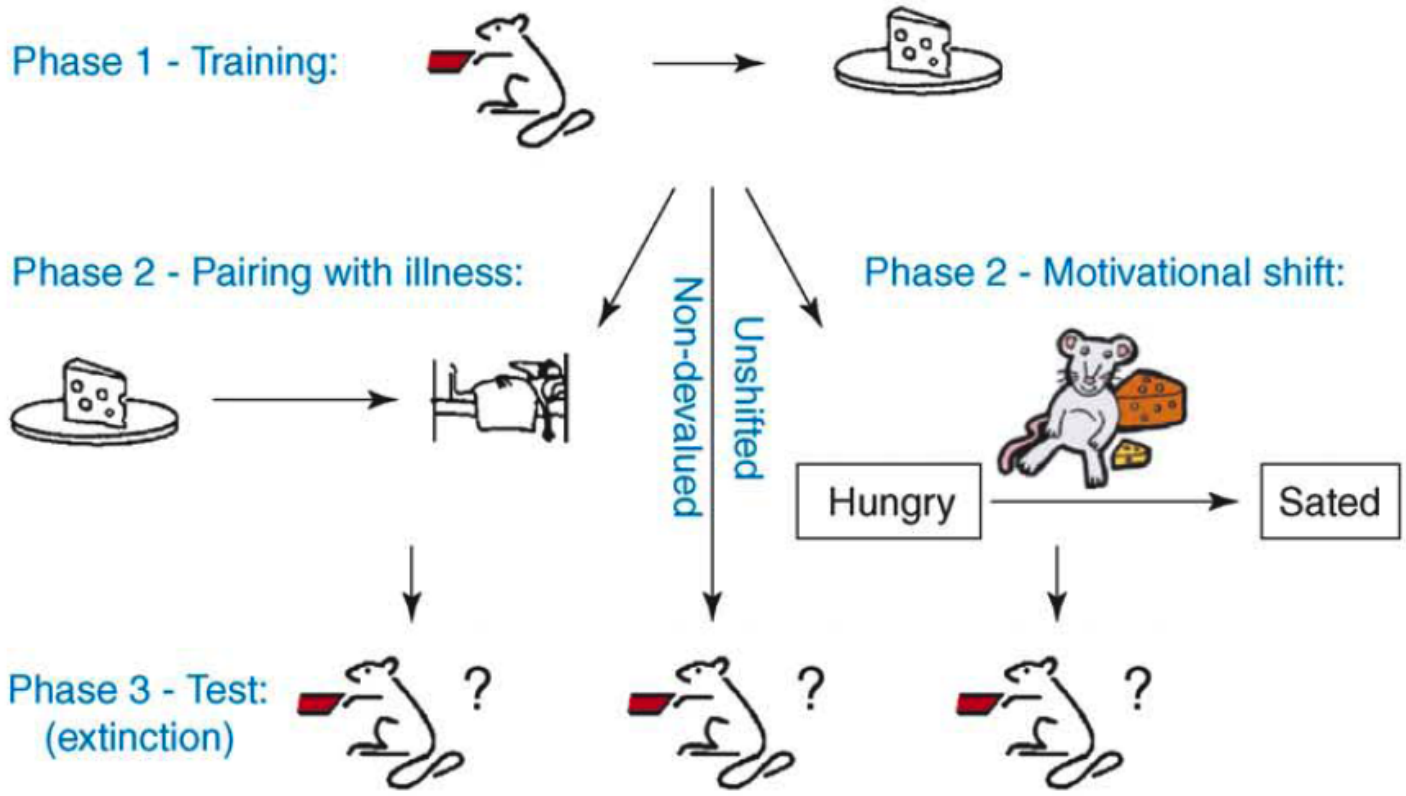
- given learnt values, behavior is a fixed SR (or state-action) mapping
- if the “goal” changes: need to re-learn values for every state in the world! all previous values are obsolete
- no general “knowledge”, only values
- no anticipation of general outcomes (s'), only of value
- no “planning”

model-free RL?

NO WAY!

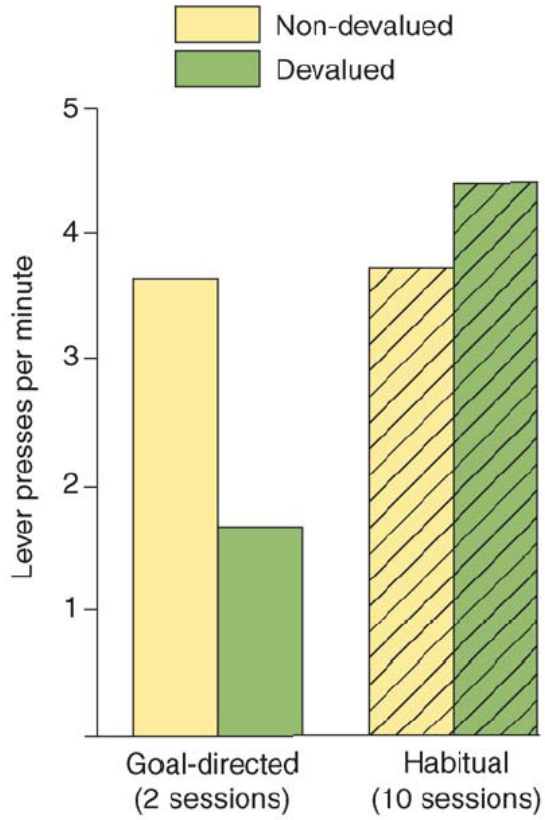
Goal-directed vs. habitual: Devaluation

[movie]



Niv, Joel & Dayan: *A normative perspective on motivation.* TICS, 10:375-381, 2006.

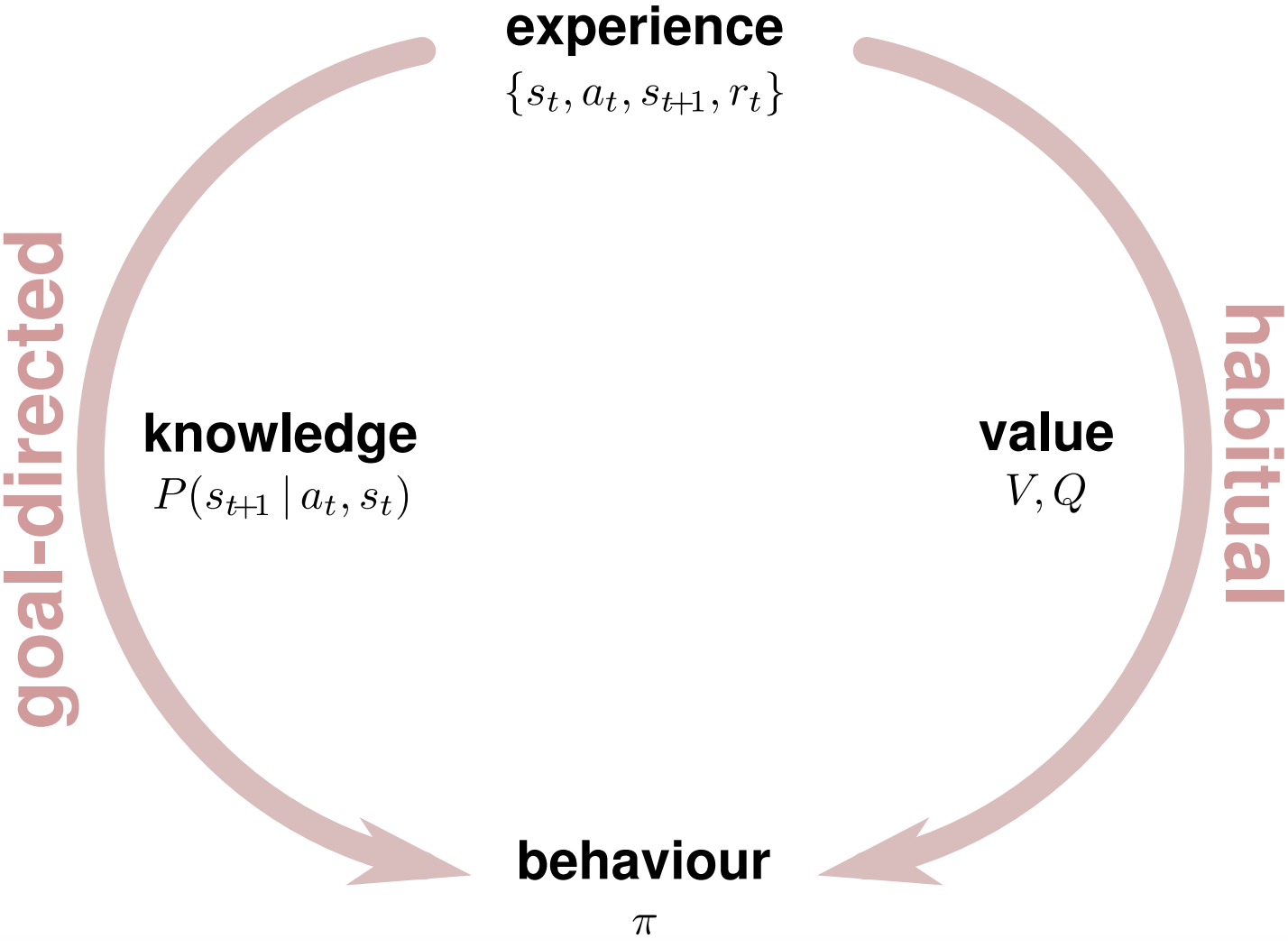
Goal-directed vs. habitual: Devaluation



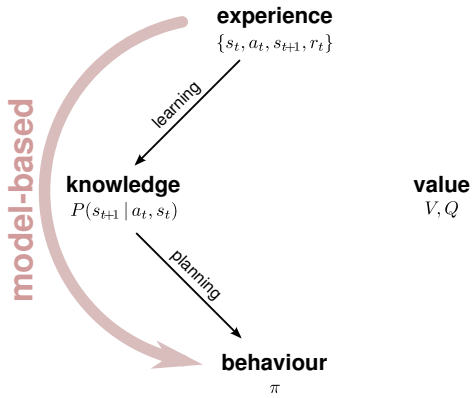
Niv, Joel & Dayan: *A normative perspective on motivation.* TICS, 10:375-381, 2006.

By definition, goal-directed behavior is performed to obtain a desired goal. Although all instrumental behavior is **instrumental** in achieving its contingent goals, it is not necessarily purposively **goal-directed**. Dickinson and Balleine [1,11] proposed that behavior is goal-directed if: (i) it is sensitive to the contingency between action and outcome, and (ii) the outcome is desired. Based on the second condition, motivational manipulations have been used to distinguish between two systems of action control: if an instrumental outcome is no longer a valued goal (for instance, food for a sated animal) and the behavior persists, it must not be goaldirected. Indeed, after moderate amounts of training, outcome revaluation brings about an appropriate change in instrumental actions (e.g. leverpressing) [43,44], but this is no longer the case for extensively trained responses ([30,31], but see [45]). That extensive training can render an instrumental action independent of the value of its consequent outcome has been regarded as the experimental parallel of the folk psychology maxim that wellperformed actions become **habitual** [9] (see Figure I).

Niv, Joel & Dayan: *A normative perspective on motivation.* TICS, 10:375-381, 2006.


experience
$\{s_t, a_t, s_{t+1}, r_t\}$
goal-directed
knowledge
$P(s_{t+1} \mid a_t, s_t)$
value
V, Q
habitual
behaviour
π

Model-based RL



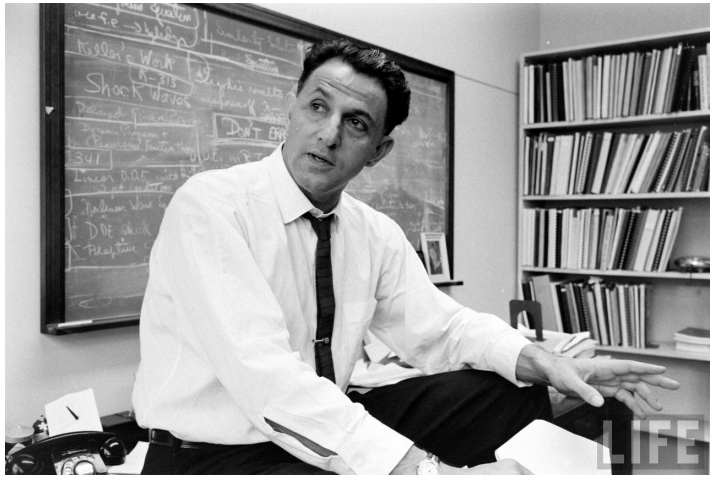
- *in this lecture:* we neglect the model learning part
  – use experience data to estimate model $\hat{P}(x'|a, x) \propto \#(x' \leftarrow x|a)$
  – for non-direct representations: Machine Learning methods

- core question:

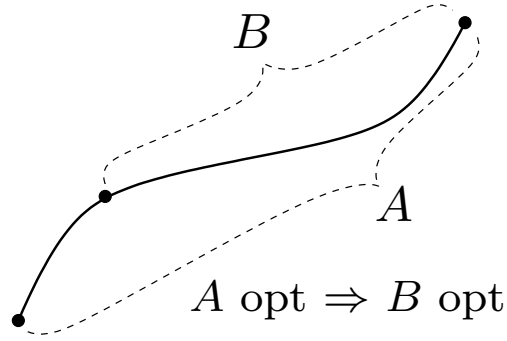
  Given an (approximate/learnt) predictive model,how choose behavior?

  What are computational principles for "planning"?

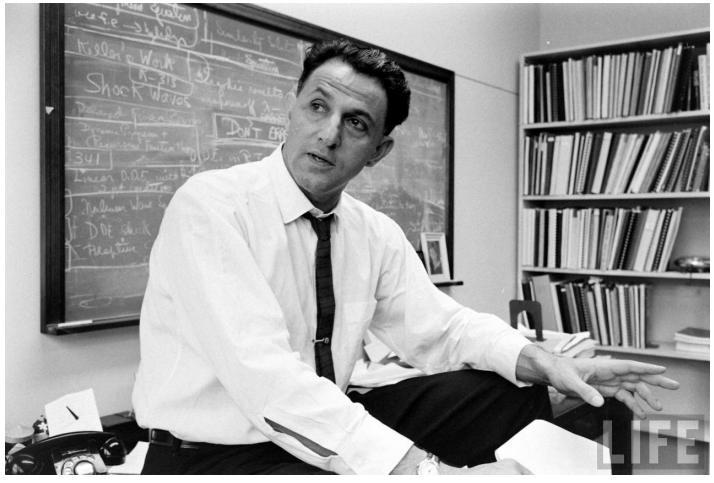
Richard E. Bellman (1920-1984)



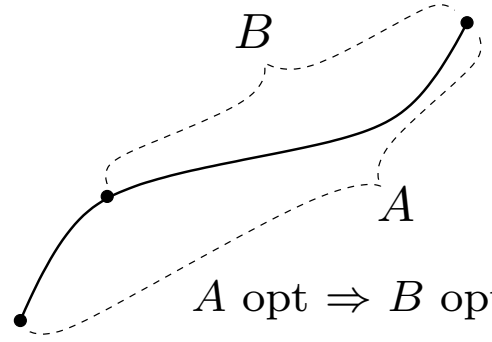
Bellman's principle of optimality



Richard E. Bellman (1920-1984)



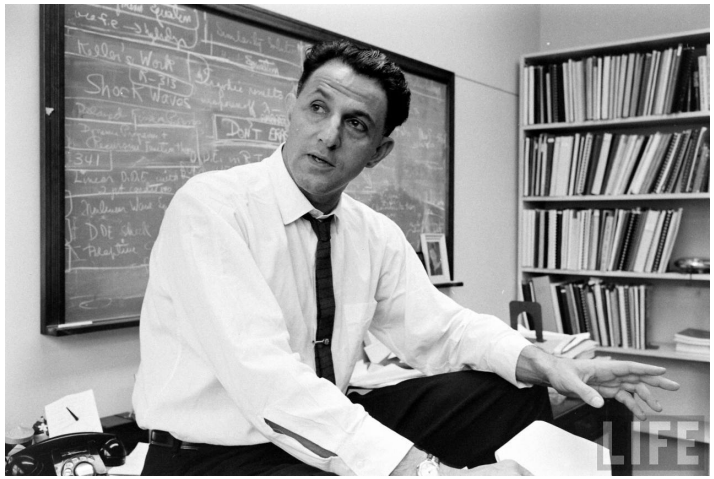
Bellman's principle of optimality



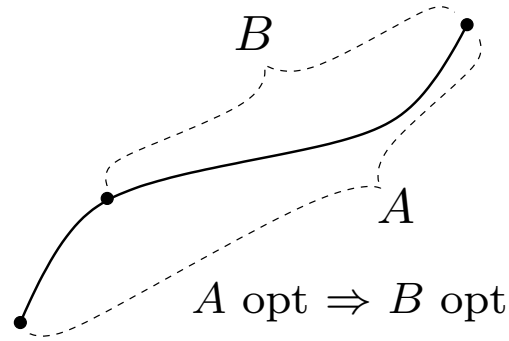
$$V(s) = \max_a \left[\mathcal{R}(a,s) + \gamma \sum_{s'} P(s' \mid a, s)\, V(s')\right]$$

$$\pi(s) = \operatorname{argmax}_a \left[\mathcal{R}(a,s) + \gamma \sum_{s'} P(s' \mid a, s)\, V(s')\right]$$

Richard E. Bellman (1920-1984)



Bellman's principle of optimality



$$V(s) = \max_a \left[\mathcal{R}(a, s) + \gamma \sum_{s'} P(s' \mid a, s)\, V(s') \right]$$

$$\pi(s) = \operatorname{argmax}_a \left[\mathcal{R}(a, s) + \gamma \sum_{s'} P(s' \mid a, s)\, V(s') \right]$$

- Rational:

  1) We can use the model (instead of experience) to compute values

  2) Then, again, derive behavior from the value function

Alternative thoughts on thought

Rick Grush (Behavioral and Brain Sciences, 2004):
The emulation theory of representation: motor control, imagery, and perception.
(20 pages + 46 pages commentary & response!)

keywords: *Kalman filters, overt & covert actions, imagery*

Alternative thoughts on thought

G. Hesslow (Trends in Cog Sciences, 2002):

Conscious thought as simulation of behaviour and perception.

A 'simulation' theory of cognitive function can be based on three assumptions about brain function.
(1) First, behaviour can be simulated by activating motor structures, as during an overt action but suppressing its execution.
(2) Second, perception can be simulated by internal activation of sensory cortex, as during normal perception of external stimuli.
(3) Third, both overt and covert actions can elicit perceptual simulation of their normal consequences.

A large body of evidence supports these assumptions. It is argued that the simulation approach can explain the relations between motor, sensory and cognitive functions and the appearance of an inner world.

Johnson & Redish (J o Neuroscience, 2007): *Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point*

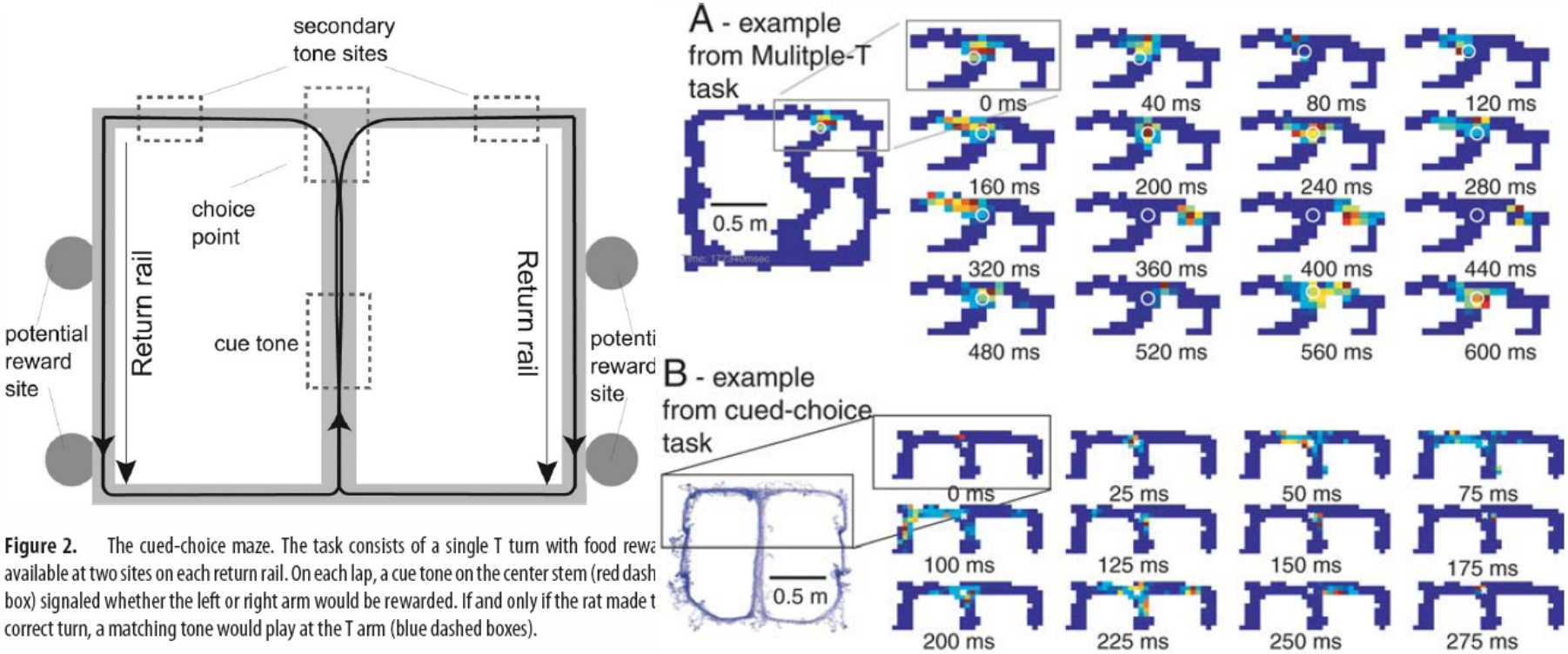


Figure 2. The cued-choice maze. The task consists of a single T turn with food rewa available at two sites on each return rail. On each lap, a cue tone on the center stem (red dash box) signaled whether the left or right arm would be rewarded. If and only if the rat made t correct turn, a matching tone would play at the T arm (blue dashed boxes).

Bottom Line

people have this idea of cognition/thinking/planning as internal simulation

Bottom Line

people have this idea of cognition/thinking/planning as internal simulation

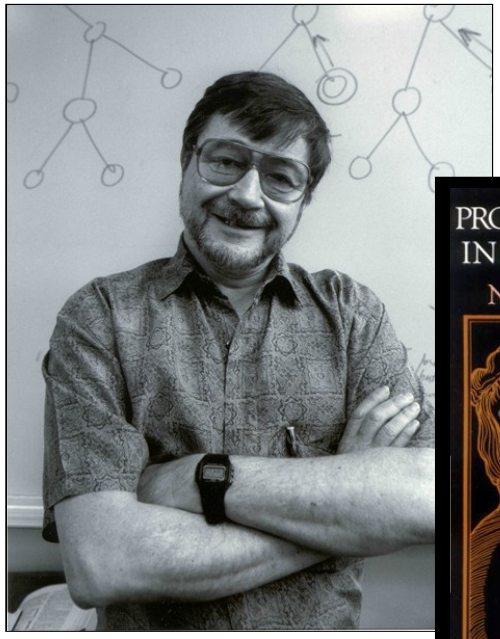
wanted:

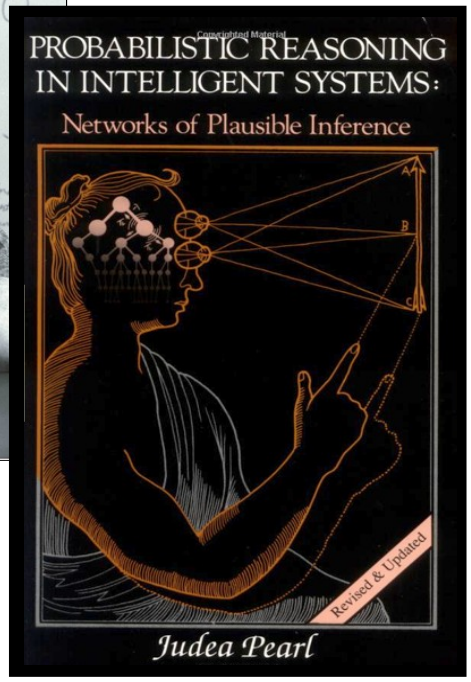
– proper math to express this

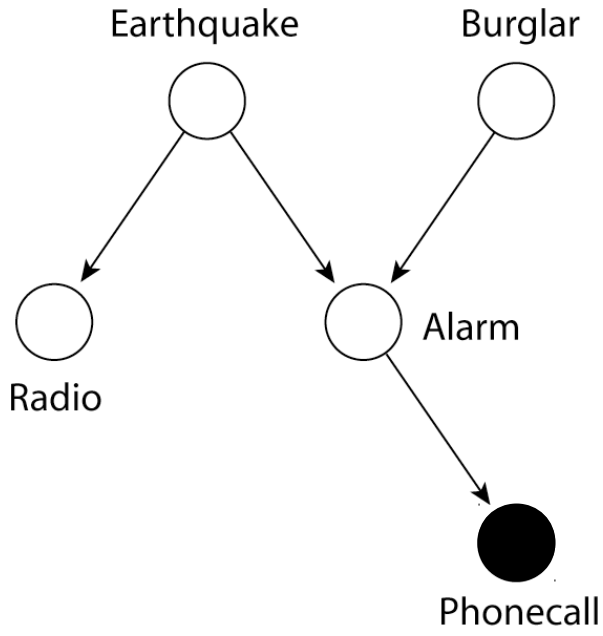
– computational perspective

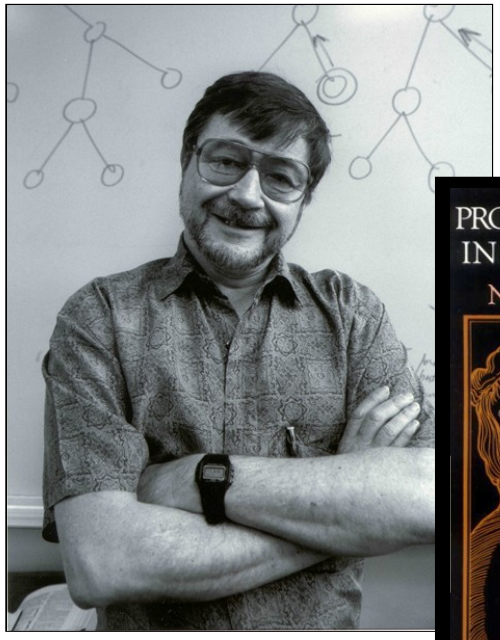
– neural perspective

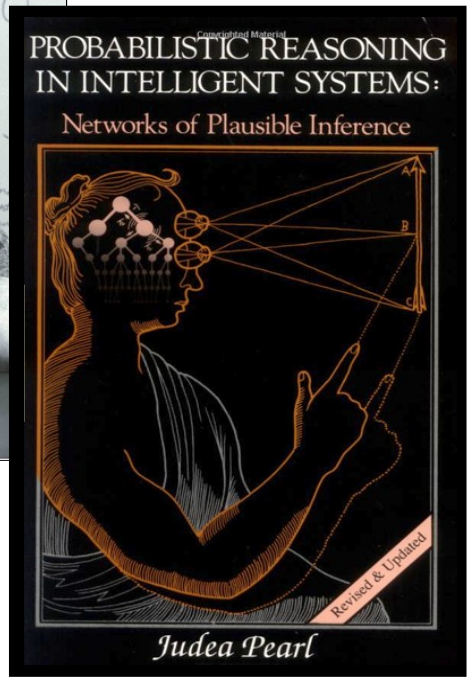
– cogsci perspective

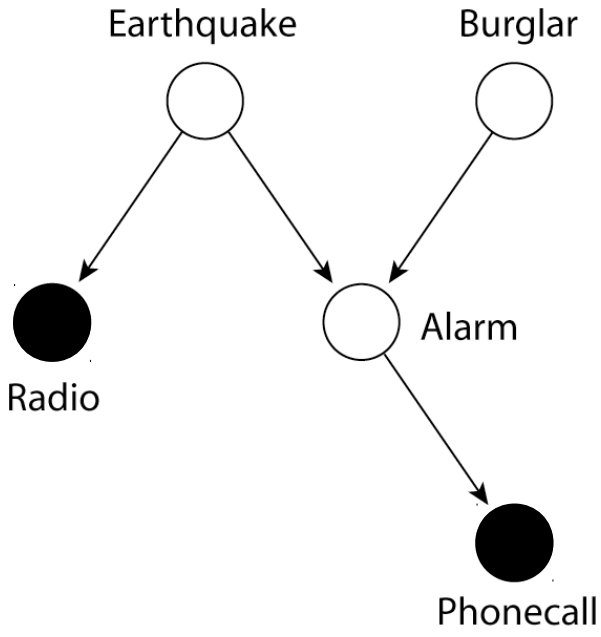


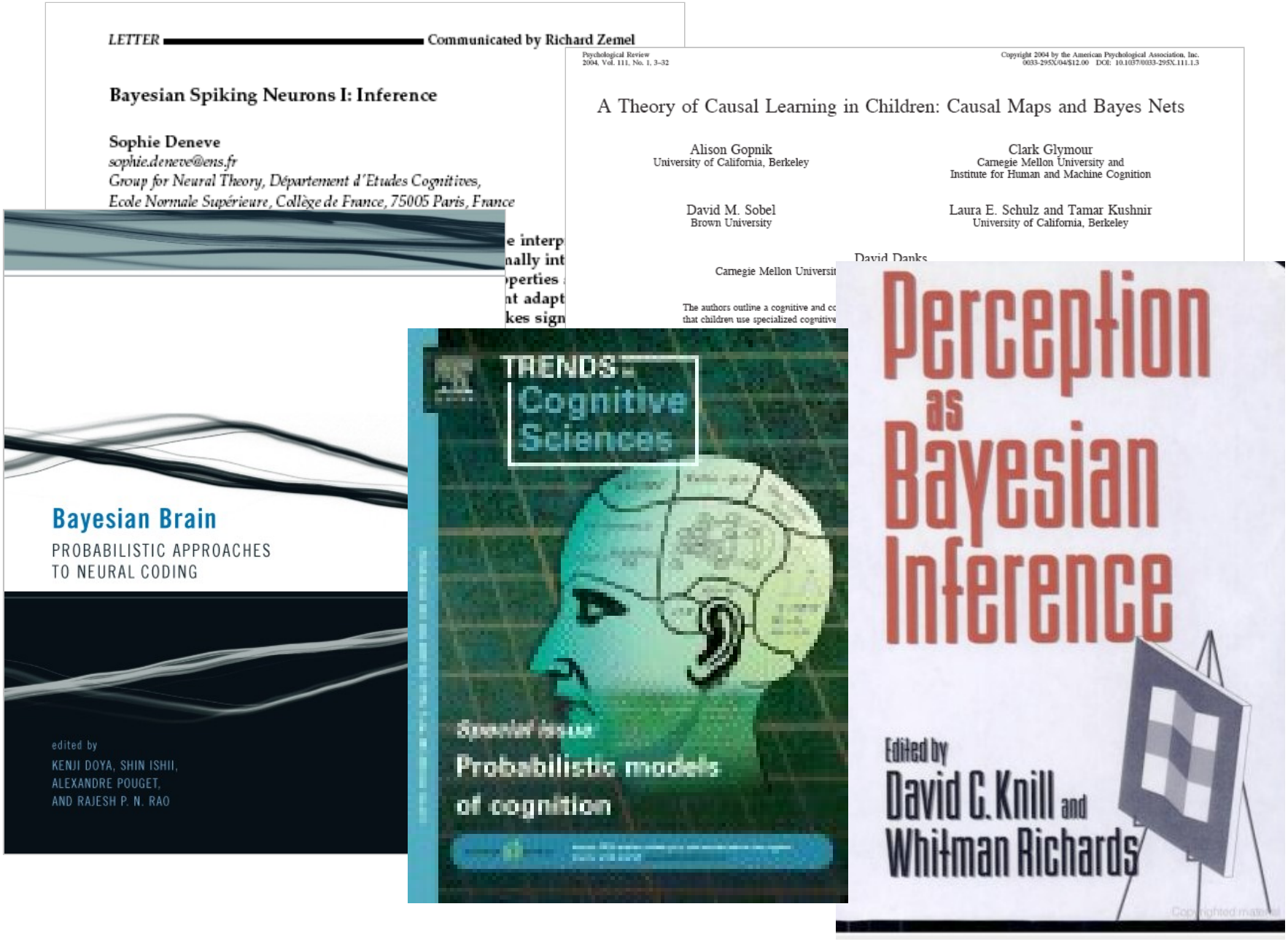

PROBABILISTIC REASONING
IN INTELLIGENT SYSTEMS:
Networks of Plausible Inference
Revised & Updated
Judea Pearl


Earthquake
Burglar
Alarm
Radio
Phonecall

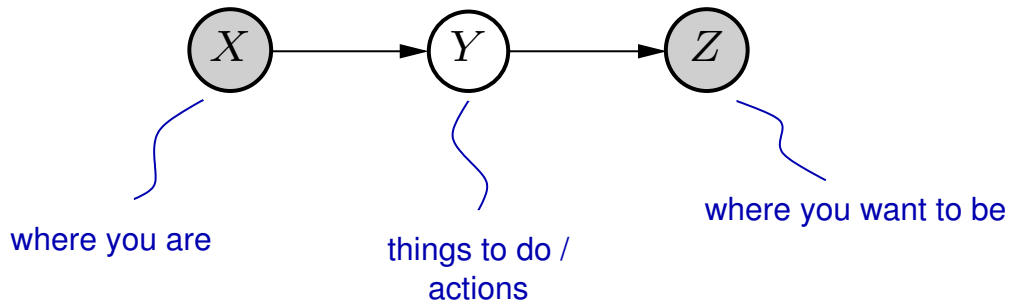



PROBABILISTIC REASONING
IN INTELLIGENT SYSTEMS:
Networks of Plausible Inference
Revised & Updated
Judea Pearl


Earthquake
Burglar
Alarm
Radio
Phonecall


LETTER
Communicated by Richard Zemel
Bayesian Spiking Neurons I: Inference
Sophie Deneve
sophie.deneve@ens.fr
Group for Neural Theory, Département d'Etudes Cognitives,
Ecole Normale Supérieure, Collège de France, 75005 Paris, France
Psychological Review
2004, Vol. 111, No. 1, 3–32
Copyright 2004 by the American Psychological Association, Inc.
0033-295X/04/$12.00 DOI: 10.1037/0033-295X.111.1.3
A Theory of Causal Learning in Children: Causal Maps and Bayes Nets
Alison Gopnik
University of California, Berkeley
Clark Glymour
Carnegie Mellon University and
Institute for Human and Machine Cognition
David M. Sobel
Brown University
Laura E. Schulz and Tamar Kushnir
University of California, Berkeley
Bayesian Brain
PROBABILISTIC APPROACHES
TO NEURAL CODING
edited by
KENJI DOYA, SHIN ISHII,
ALEXANDRE POUGET,
AND RAJESH P. N. RAO
TRENDS in Cognitive Sciences
Special issue
Probabilistic models of cognition
Perception as Bayesian Inference
Edited by
David C. Knill and
Whitman Richards

Planning by Inference



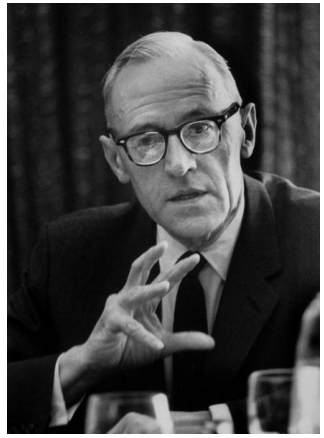
– *What is the posterior over intermediate actions?*

- This implies the downfall of the notion of reward!
  Rewards are replaced by “desired future stimuli”

We condition our future and infer our present!

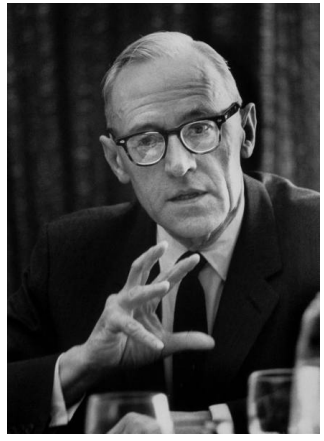
von Neumann
(1903-1957)

Morgenstern
(1902-1977)



von Neumann
(1903-1957)

Morgenstern
(1902-1977)

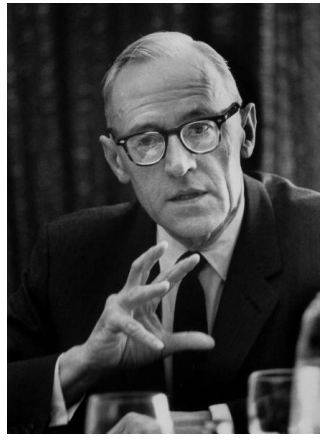


Shachter
(Stanford)

Cooper
(Pittsburgh)

von Neumann
(1903-1957)

Morgenstern
(1902-1977)



Shachter
(Stanford)

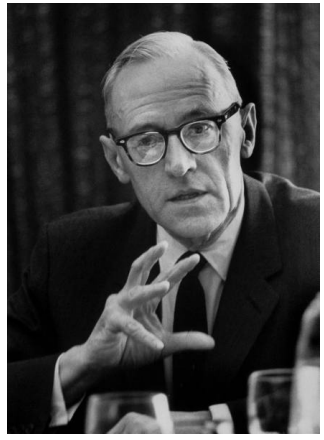
Cooper
(Pittsburgh)

utility/reward/success

$\sim$ *just another state (random) variable*

von Neumann
(1903-1957)

Morgenstern
(1902-1977)



utility/reward/success

$\sim$ *just another state (random) variable*

Shachter
(Stanford)

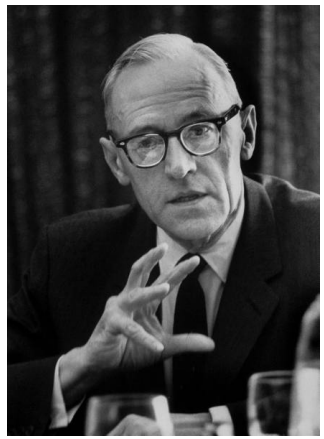
Cooper
(Pittsburgh)

Shachter/Cooper/Peot (1988):

inference $P(\text{behavior} \mid R = 1)$

von Neumann
(1903-1957)

Morgenstern
(1902-1977)



utility/reward/success

$\sim$ *just another state (random) variable*

Shachter
(Stanford)

Cooper
(Pittsburgh)

Shachter/Cooper/Peot (1988):

inference $P(\text{behavior} \mid R = 1)$

– I think: idea was sleeping for 15 years

– should have had great influence on thinking about behavior!

– a group of MLs independently rediscovered this recently

history

– inference in influence diagrams: Cooper (1988), Shachter & Peot (1992)
no sequential decisions, no discounting, finite horizon

history

– inference in influence diagrams: Cooper (1988), Shachter & Peot (1992)
  no sequential decisions, no discounting, finite horizon

– Dayan & Hinton (1997): neuroscience, immediate reward case only
– Littman et al. (2001): complexity equiv. between planning and inference
– Attias (2003): “Planning by Probabilistic Inference”
  non-optimal policy, no arbitrary rewards, prefixed finite time
– Raiko (2005): optimistic inference control
– Bui et al. (2002): policy recognition/goal inference
– Verma & Rao (2006): maximal probable explanation (MPE)

history

– inference in influence diagrams: Cooper (1988), Shachter & Peot (1992)
  no sequential decisions, no discounting, finite horizon

– Dayan & Hinton (1997): neuroscience, immediate reward case only
– Littman et al. (2001): complexity equiv. between planning and inference
– Attias (2003): "Planning by Probabilistic Inference"
  non-optimal policy, no arbitrary rewards, prefixed finite time
– Raiko (2005): optimistic inference control
– Bui et al. (2002): policy recognition/goal inference
– Verma & Rao (2006): maximal probable explanation (MPE)

– Toussaint & Storkey (ICML 2006): EM, optimality in MDPs
– Hoffman et al. (NIPS 2007): MCMC methods for policy learning
– Toussaint, Charlin, Poupart (UAI 2008): hierarchical POMDPs
– Kober & Peters (NIPS 2008): Motor primtive learning using EM
– Botvinick & An (NIPS 2008): prefrontal cortext
– Vlassis & Toussaint (ICML 2009): model-free RL

Theoretical perspective

Optimality w.r.t. reward can be also be achieved with

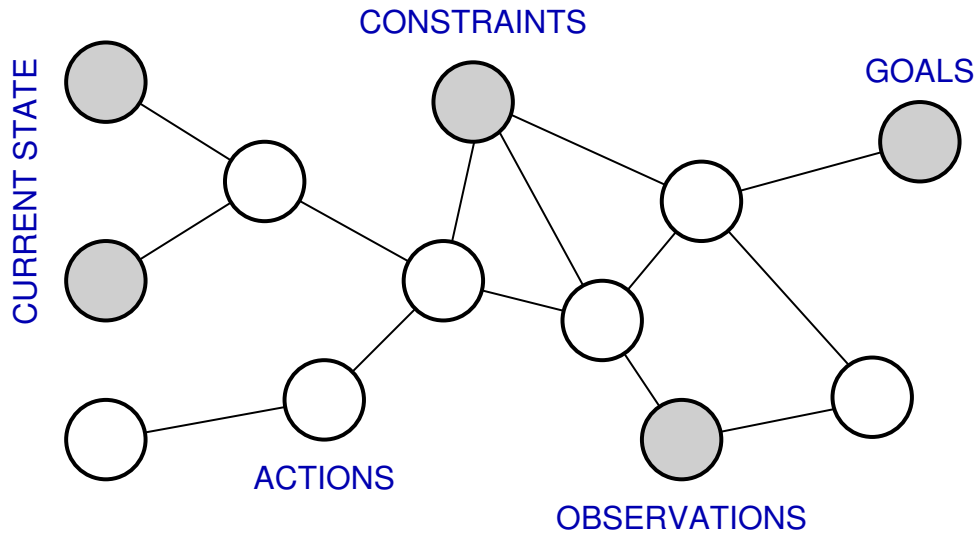
Planning by Inference

(condition your future on observing lots of rewards...)

Toussaint, Storkey (ICML 2006): *Probabilistic inference for solving discrete and continuous state Markov Decision Processes*

But there is no need for rewards,

the setting is more general...



works for *networks* of goals, constraints, observations, etc.

– doesn't distinguish between sensor/motor, perception/action

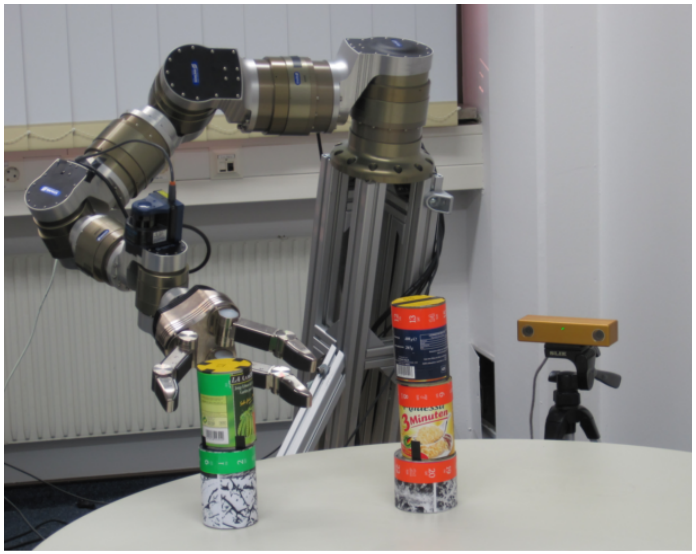
– we got rid of notion of *state* as one big variable

Computational perspective

- problem of planning becomes problem of inference!
  belief propagation, variational inference, sampling
  workd on distributed, hybrid representations etc

- inference over temporal processes usually involves forward-backward sweeps
  (recall Johnson & Redish's rat experiment)

- the specific algorithms are typically specialized:
  – hierarchical POMDPs
  – Stochastic Optimal Control
  – Relational Planning



- hardware:
  - Schunk arm LWA (7DoF)
  - Schunk hand SDH (7DoF)
  - tactile sensor arrays
  - vision (Bumblebee stereo camera)
- 14 joints, dynamic $\rightarrow x_t \in \mathbb{R}^{28}$

- PRADA to plan on the stochastic relational level (Tobias Lang)
- AICO to generate fluent reach-and-pre-grasp trajectories

  we condition on:

  – no collisions, no limits along the whole trajectory

  – final endeffector (center of palm) position = object position

  – final finger-surface distance = 3cm

  – final finger normals are opposing

Toussaint et al. (ICRA 2010): *Integrated motor control, planning, grasping and high-level reasoning in a blocks world using probabilistic inference*

Neural perspective

Matthew Botvinick

(Princeton)

Neural perspective

Matthew Botvinick
(Princeton)

Goal-directed decision making in prefrontal cortex: A computational framework (NIPS 2008)

"We take three empirically motivated points as founding premises: (1) Neurons in dorsolateral prefrontal cortex represent action policies, (2) Neurons in orbitofrontal cortex represent rewards, and (3) Neural computation, across domains, can be appropriately understood as performing structured probabilistic inference. [...]"

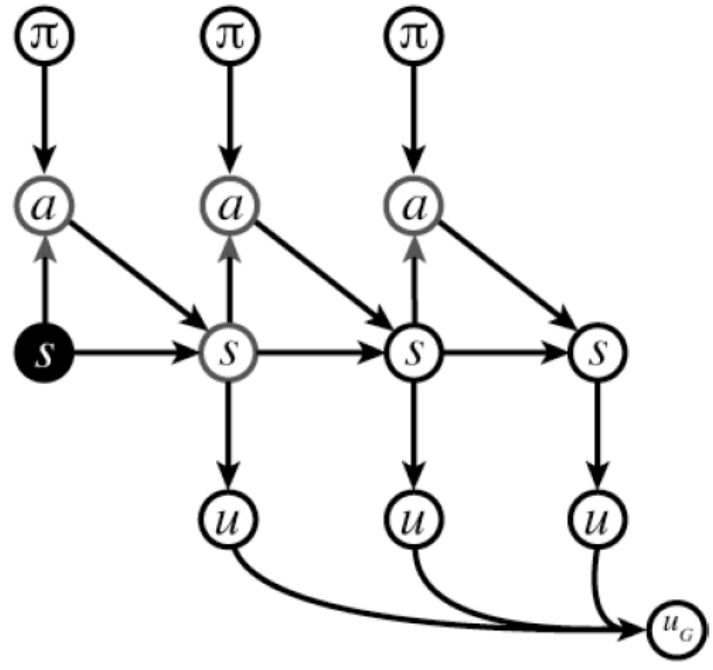
... essentially proposes Planning by Inference

Policy System
Dorsolateral PFC
Premotor, VLPFC, BA 10

Action System
Premotor cortex, SMA

State Projection System
Medial temporal cortex
Medial frontal / parietal cortex
Caudate nucleus

Reward System
Orbitofrontal cortex
Basolateral amygdala



Karl Friston (UCL)

Karl J. Friston, Jean Daunizeau, James Kilner, Stefan J. Kiebel: *Action and behavior: a free-energy formulation*, Biol Cyb.

“minimize sensory prediction error (free-energy) through action”

Probabilistic Inference and Free Energy Minimization are equivalent (in the sense of Kullback Leibler divergence minimization)

Inference in neural circuits?

Inference in neural circuits?

- message passing in Markov processes:

$$\hat{\alpha}_i^{t+1} = P(Y \mid x_{t+1}=i) \sum\nolimits_j P(x_{t+1}=i \mid x_t=j)\ \alpha_j^t$$
$$\alpha_i^{t+1} = \hat{\alpha}_i^{t+1} / \sum\nolimits_j \hat{\alpha}_j^{t+1}$$
$$\log \hat{\alpha}_i^{t+1} = \log P(Y \mid x_{t+1}=i) + \log \sum\nolimits_j P_{ij}\ \alpha_j^t$$

- leaky integrator in a recurrent NN:

$$v_i(t+1) = v_i(t) + \epsilon[-v_i(t) + w_i^\top I(t) + \sum\nolimits_j u_{ij} v_j(t)]$$
$$=: \epsilon w_i^\top I(t) + \sum\nolimits_j U_{ij} v_j(t)$$

- equivalent if

$$v_i(t+1) = \log \hat{\alpha}_i^{t+1}$$
$$\epsilon w_i^\top I(t) = \log P(Y \mid x_{t+1}=i)$$
$$\sum\nolimits_j U_{ij} v_j(t) = \log \sum\nolimits_j P_{ij} \alpha_j^t$$

Rajesh P. N. Rao: *Neural Models of Bayesian Belief Propagation*, 2006.

Inference in neural circuits?

- message passing in Markov processes:

$$\hat{\alpha}_i^{t+1} = P(Y \mid x_{t+1}=i) \sum\nolimits_j P(x_{t+1}=i \mid x_t=j)\ \alpha_j^t$$
$$\alpha_i^{t+1} = \hat{\alpha}_i^{t+1} / \sum\nolimits_j \hat{\alpha}_j^{t+1}$$
$$\log \hat{\alpha}_i^{t+1} = \log P(Y \mid x_{t+1}=i) + \log \sum\nolimits_j P_{ij}\ \alpha_j^t$$

- leaky integrator in a recurrent NN:

$$v_i(t+1) = v_i(t) + \epsilon[-v_i(t) + w_i^\top I(t) + \sum\nolimits_j u_{ij} v_j(t)]$$
$$=: f(\epsilon w_i^\top I(t)) + g(\sum\nolimits_j U_{ij} v_j(t))$$

- equivalent if

$$v_i(t+1) = \log \hat{\alpha}_i^{t+1}$$
$$f(\epsilon w_i^\top I(t)) = \log P(Y \mid x_{t+1}=i)$$
$$g(\sum\nolimits_j U_{ij} v_j(t)) = \log \sum\nolimits_j P_{ij} \alpha_j^t$$

Rajesh P. N. Rao: *Neural Models of Bayesian Belief Propagation*, 2006.

Bayesian Brain: Probabilistic Approaches to Neural Coding. K. Doya, S. Ishii, A. Pouget, RPN. Rao (editors), MIT Press (2007)

Hierarchical Bayesian Inference in Networks of Spiking Neurons. Rajesh P. N. Rao (NIPS 2004)

The Neurodynamics of Belief Propagation on Binary Markov Random Fields. T. Ott, R. Stoop (NIPS 2006)

current work by Wolfgang Maas ...

Johnson & Redish (J o Neuroscience, 2007): *Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point*

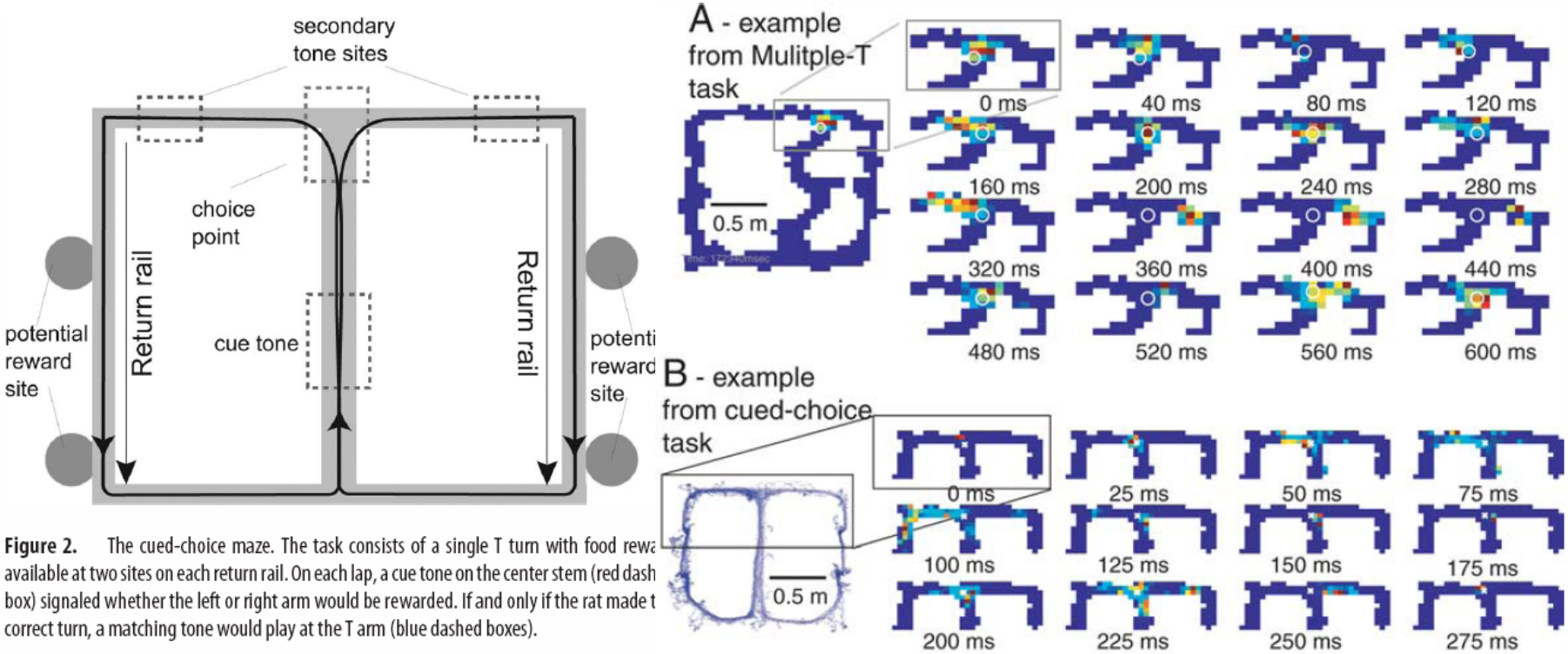


Figure 2. The cued-choice maze. The task consists of a single T turn with food rewa available at two sites on each return rail. On each lap, a cue tone on the center stem (red dash box) signaled whether the left or right arm would be rewarded. If and only if the rat made t correct turn, a matching tone would play at the T arm (blue dashed boxes).

speculative . . .

- What neuroscientific experiments would discriminate between computational hypotheses?

speculative . . .

- What neuroscientific experiments would discriminate between computational hypotheses?

  If taking *"We condition our future and infer our present"* literal

  $$\text{having a goal} \quad \leftrightarrow \quad \text{seeing the goal}$$

  Model-free view (value function) would not predict this

CogSci perspective

speculative . . .

CogSci perspective

speculative . . .

- What psychophysical experiments would discriminate between computational hypotheses?

CogSci perspective

speculative . . .

- What psychophysical experiments would discriminate between computational hypotheses?

  Testing optimality? No!

  (Many computational paradigms can generate optimal behaviors.)

CogSci perspective

speculative . . .

- What psychophysical experiments would discriminate between computational hypotheses?

  Testing optimality? No!

  (Many computational paradigms can generate optimal behaviors.)

  Testing response time, failures, results under pressure, non-optimal results!

response time:

– dominating model in psychology: drift-diffusion models

– closely related to sampling

response time:

– dominating model in psychology: drift-diffusion models

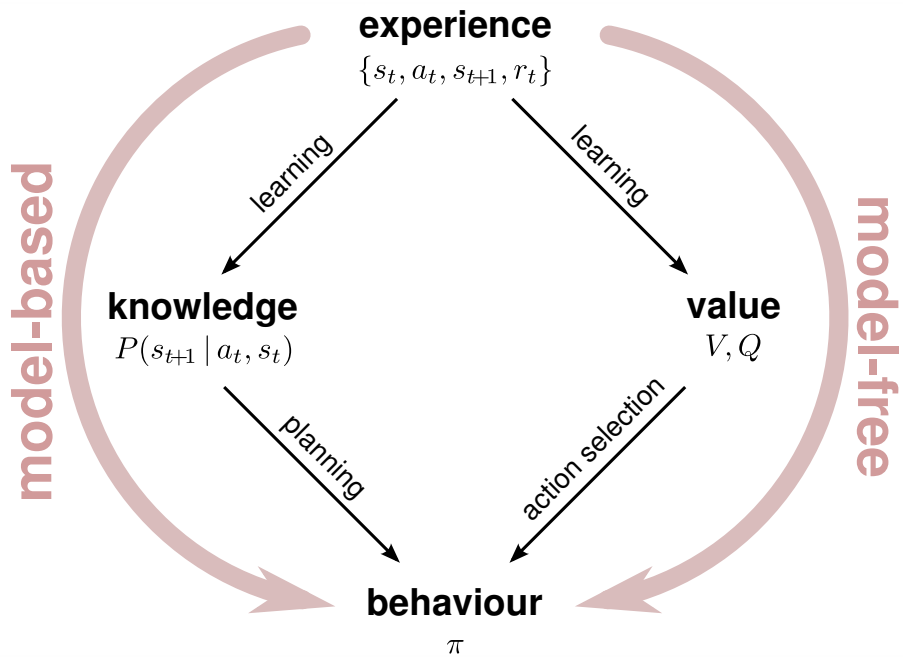
– closely related to sampling

hypothesis:

sampling-based inference for planning

$\leftrightarrow$

response time models for goal-directed behavior



model-free RL $\leftrightarrow$ habitual $\leftrightarrow$ TD learning $\leftrightarrow$ Dopamine

model-based RL $\leftrightarrow$ goal-directed $\leftrightarrow$ Planning by Inf. $\leftrightarrow$ neural inf.

[movie]