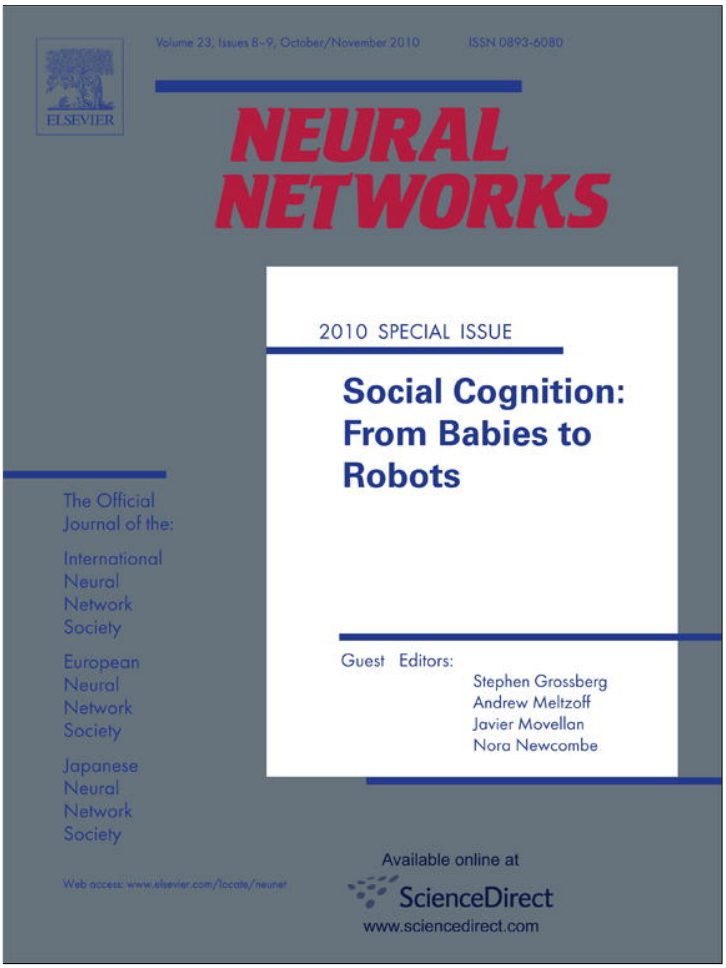

Neural Networks 23 (2010) 973–984

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2010 Special Issue

Detecting contingencies: An infomax approach

Nicholas J. Butko *, Javier R. Movellan

University of California, San Diego, United States

ARTICLE INFO

Keywords:
Active learning
Real-time learning
Information maximization
Contingency detection
Social robotics
Control theory

ABSTRACT

The ability to detect social contingencies plays an important role in the social and emotional development of infants. Analyzing this problem from a computational perspective may provide important clues for understanding social development, as well as for the synthesis of social behavior in robots. In this paper, we show that the turn-taking behaviors observed in infants during contingency detection situations are tuned to optimally gather information as to whether a person is responsive to them. We show that simple reinforcement learning mechanisms can explain how infants acquire these efficient contingency detection schemas. The key is to use the reduction of uncertainty (information gain) as a reward signal. The result is an interesting form of learning in which the learner rewards itself for conducting actions that help reduce its own sense of uncertainty. This paper illustrates the possibilities of an emerging area of computer science and engineering that focuses on the computational understanding of human behavior and on its synthesis in robots. We believe that the theory of stochastic optimal control will play a key role providing a formal mathematical foundation for this newly emerging discipline.

1. Introduction

Peter picks up the phone: "Hello, this is Peter", he says. A voice responds, "Arratsalde on... Zer moduz". Surprised, Peter repeats, "Hello, this is Peter". The voice responds, "Euskaraz badakizu?" Peter says, "I think you are calling the wrong number. Who are you trying to reach?" The voice responds, "Ez dut ulertzen. Astiro-astiro hitz egin mesedz". Peter did not understand a single word, but he had the distinct impression that there was a person trying to communicate with him at the other end of the line. It did not feel at all like a pre-recorded message.

Infants face situations like this very early in their lives. They do not understand human language, but they still need to identify what entities are responsive to them and when they are so. Developmental psychologists refer to this ability to identify responsive entities as "contingency detection", "contingency analysis", "contingency perception", and "contingency learning".

There is a large body of evidence suggesting that the ability to detect contingencies plays a crucial role in the social and emotional development of infants (Bigelow, 1999; Bahrick & Watson, 1985; Watson, 1972, 1979, 1985). For example, it has been hypothesized that infants use contingency, not appearance, as the main cue to detect conspecifics. The appearance of human beings becomes special to infants because they can generate contingencies. This point of view traces back to an experiment conducted by John Watson in 1972. In this experiment, 2-month-old infants learned to move their heads to activate a mobile located above their cribs (Watson, 1972). Each infant in the experimental group was presented with a mobile that rotated in response to the motion of her head. For the infants in the control group, the mobile moved in a pre-recorded, non-contingent manner. After four daily 10 min sessions, and an average of 200 total responses, there was evidence that the infants in the experimental group had learned that they could control the mobile. At the same time, these infants displayed a number of powerful social responses towards the mobile, including vigorous cooing and smiling. Essentially the mobile began functioning as a "social stimulus". Watson hypothesized that contingency was being used by these infants as a cue to define and identify caregivers.

Movellan and Watson (1987, 2002) conducted a similar experiment with 10-month-old infants. Infants were seated in front of a robot that did not look particularly human. The "head" of the robot was a rectangular prism whose sides contained geometric patterns (see left side of Fig. 1). The robot could make sounds and turn its head to the right or left. Infants were randomly assigned to an experimental group or a matched control group. In the experimental group, the robot produced sounds in response to the infants' vocalizations. In the control group, the robot reproduced the same responses that had been recorded in the matched experimental session. In this way, infants in the control group experienced exactly the same robot activity, except that it was pre-recorded and not responsive to them. After a few minutes of exposure to the robot, many infants in the experimental group were treating the robot as if it were a social agent: they produced 5 times more vocalizations than the infants in the control

* Corresponding author. Tel.: +1 858 3377117; fax: +1 858 3377117.
E-mail address: nbutko@ucsd.edu (N.J. Butko).

0893-6080/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2010.09.001

Fig. 1. Left: schematic of the robot head used in Movellan and Watson (1987). Right: Baby-9. The image of the robot is seen reflected on a mirror positioned behind the baby.

group, and they followed the "line of regard" of the robot when it rotated (Movellan & Watson, 1987, 2002). Similar results were later replicated with 12-month-old children (Johnson, Slaughter, & Carey, 1998).

Particularly striking was the quality of interactions that were observed in some infants in the experimental group: (1) Their vocalizations toward the robot appeared to be like questions. Each vocalization was followed by 5–7 s of silence, during which the infants seemed to be actively waiting for an answer from the robot. (2) After a few such vocalizations and less than a minute into the experiment, most observers report that these infants know that the robot is responding to them.

The video of one such baby, hereafter named "Baby-9," will be the focus of this document. This video is available at doi:10.1016/j.neunet.2010.09.001 and is an essential companion to this document. The reader is recommended to watch this video to better understand the focus of this paper. Most people that watch the video report that Baby-9 has clearly detected the responsiveness of the robot. Many of these additionally indicate that Baby-9 is actively querying the robot, as if questioning whether or not it is responsive.

Challenge problems: understanding the pattern of behavior that Baby-9 exhibited poses theoretical challenges with important consequences for the scientific study of social development in infants:

1. What does it mean to "ask questions" for an organism like Baby-9 that does not have language?
2. Was it smart for Baby-9 to schedule his vocalizations in the way that he did?
3. Was it smart for him to decide within a few responses and less than a minute into the experiment that the robot was responsive?
4. What mechanisms can explain the transition from the relatively slow learning that Watson observed in 2-month-old infants to the very fast and active learning that was observed in 10-month-old infants like Baby-9?

In this paper, we explore a computational approach to these theoretical questions based on the framework of stochastic optimal control. Originally developed by engineers to control complex systems like airplanes and industrial robots, stochastic optimal control is giving behavioral scientists a unifying theory to describe diverse human skills such as reaching, walking, eye-movements, and concept learning (Bertsekas & Shreve, 1996; Butko & Movellan, 2008, 2009, 2010; Nelson & Movellan, 2001; Nelson, Tenenbaum, & Movellan, 2001). We propose that the same framework can be used to understand the development of social interaction. In particular, the behavior observed in Baby-9's video can be seen as a sensory–motor schema optimized for gathering information as to whether or not a social contingency is present.

In the paper we show how social skills, like the ones observed in Baby-9, could be acquired using standard reinforcement learning mechanisms. The key for this to happen is to use information as an intrinsic reward. This opens the possibility that the same mechanisms that are used to learn how to reach, walk, and look, could also be used to acquire social skills, including the development of symbolic communication.

A long term goal of this paper is to illustrate how stochastic optimal control may be used to provide a computational basis for the study of human development. The approach provides a modern alternative to behaviorist approaches that were popular in the first half of the 20th century, and to cognitive/mentalist approaches that dominated in the second half. We aim for the approach illustrated in this document to provide a computational basis to help bridge the study of the brain, the study of development, and the synthesis of intelligent behavior in robots.

2. Stochastic optimal control

Due to the inherent variability of situations that organisms encounter through their lives, biological motion can seldom rely on a predetermined sequence of actions. Instead, the behavior of organisms is more like a dance with the environment, in which sensory information is continuously polled to generate actions that are tuned to the current state of the world. Influential developmental psychologists, such as Piaget, have long argued that these sensory–motor schema provide the primordial conditions out of which high-level cognitive processes develop.

Control theory is a rigorous mathematical formalism for analyzing the sensory–motor dance between complex systems and the environment. Its focus is solving the problem of how to map sensory information into motor commands to generate intelligent behavior in real time. To give the reader a better intuition for the control theory formalism, we present a simple example. The point of this example is to illustrate the different elements of the control theory formalism. Refer to the Appendix A, for information on mathematical notation and conventions.

Simple control theory example—reaching: consider a robot who is trying to reach for an object as quickly as possible, while using as little energy as possible. To analyze this scenario in the language of control theory, we must specify the relevant states x_t that the robot can encounter, actions u_t that the robot can take to affect the state, observations y_t that the robot can use to get feedback about its progress, and the goal ρ that the robot is trying to achieve. In each of these, the momentary nature of the dance with the environment is captured by the subscript t, denoting that each element can and does constantly change.

For this problem, the relevant state x_t consists of the current angles between each of the robot's joints. The robot affects these angles by applying voltages u_t to each of its motors. The relationship between voltages and changing joint angles is captured in the world dynamics, also known as system dynamics, given by the electro-mechanic equations of motion. This is defined by a probability distribution $p(x_{t+1} \mid x_t, u_t)$ that specifies probable next states x_{t+1} given current states x_t and actions u_t. By expressing this relationship as a probability distribution, the robot can express the natural variability in the voltages it sends, as well as unpredictable external perturbations, such as people grabbing its arm.

The robot gets feedback y_t about its progress from sensors, such as encoders that measure the angle at each joint. The sensor model $p(y_{t+1} \mid x_{t+1})$ describes the encoders' readings given particular joint configurations.

The joint angles x_t determine the position p_t of the robot hand in 3D Euclidean space. The robot's goal of touching a target at a position p^* can be specified using a reward function that measures the Euclidean distance between the current position of hand and the desired posture. In addition, we could penalize actions that consume too much energy. For example, the reward could take the

following form:

$$r_t = -\|p_t - p^*\|^2 - k\|u_t\|^2 \quad (1)$$

where $k \geq 0$ is a constant that penalizes for using too much energy.

Given such a problem specification, the theory of stochastic optimal control provides algorithms to find optimal "control laws". These are also known as "policies", or simply "controllers". Controllers are the technical equivalent of the sensory–motor schemas that Piaget discussed. Formally, a control policy c is a collection of functions $c = (c_1, c_2, c_3 \ldots)$ indexed by time t. Each function c_t maps the history of data H_t available to the robot to an action U_t to be taken by the robot:

$$U_t = c_t(H_t). \quad (2)$$

The information history H_t consists of everything the robot has seen and done prior to taking an action at time t. This includes the entire history of actions $U_1 \cdots U_{t-1}$ and the entire history of sensor values $Y_1, \ldots, Y_t$, i.e.,

$$H_t = (U_1, \ldots, U_{t-1}, Y_1, \ldots Y_t). \quad (3)$$

Stochastic optimal control is essentially a computational theory of intentional, goal oriented behavior. The goals are specified using a reward variable R_t that represents the desirability of states and actions at particular points in time. The overall goal of the controller is typically expressed as a weighted sum of the expected accumulation of future rewards:

$$\rho(c) = \sum_{t=1}^{\tau} \alpha_t E[R_t \mid c] \quad (4)$$

where τ is the temporal horizon, or terminal time. Controllers are evaluated in terms of the expected reward gathered before the terminal time. Depending on the situation, this terminal time can be finite, or infinite. The α_t terms are non-negative constants that modulate the relative importance of rewards at different points in time. Stochastic optimal control considers the problem of finding control policies c that optimize the goal function $\rho(c)$.

Stochastic optimal control has been traditionally applied to optimization of physical goals (e.g., maintaining a motor's velocity under variable loads, regulating a room's temperature, and making smart weapons). In this document we show how the same approach also illuminates the development of social behavior from a computational point of view.

3. Formalizing the contingency detection problem

In order to analyze the contingency problem within the stochastic optimal control framework, we must formalize it with the same elements as the motor control problem described above: states, actions, observations, system dynamics, sensor models, and goal. Our formalization was inspired by John Watson's contingency detection model (Watson, 1985), in which background noise and responsive caregivers are modeled as Poisson processes. While Watson focused on the inference problem, i.e., the development of algorithms to infer the presence or absence of contingency given a history of sensory–motor experiences h_t, we focus on the control problem, i.e., how to schedule behaviors in real time to ensure that sensory–motor experiences h_t are as informative as possible in a limited period of time. We will investigate the problem of detecting social contingency from the point of view of a bare-bones baby robot (see Fig. 2). This idealized baby robot has a single binary sensor and a single binary actuator. The sensor tells the robot whether a sound is present, and the actuator produces vocalizations. There will be two players: (1) a *social agent* that plays the role of the caregiver, and (2) a *baby robot* that plays the role of the infant. The agent and robot are situated in an environment with random background activity. When the social agent is present, she responds to the sounds produced by the baby robot, introducing a

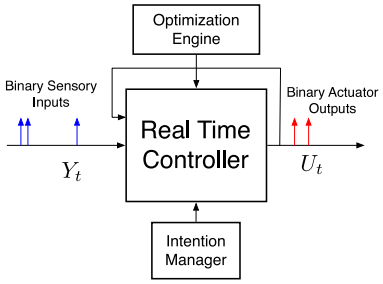

Fig. 2. A bare-bones social robot.

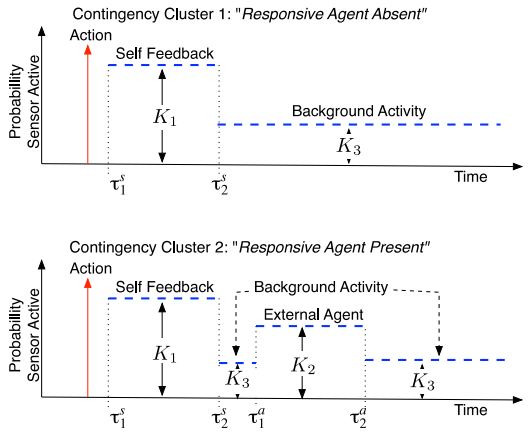

Fig. 3. Illustration of two contingency clusters produced by the model. The variable S indicates which of the two clusters is active in the current situation.

contingency between the robot's actuator and the robot's sensor. Our goal is to find an optimal control policy for the baby robot to detect, as efficiently and accurately as possible, whether or not such a contingency is present and, by extension, whether the social agent is present. While at first sight this may appear to be a simple problem, the following complications need to be considered:

- Self-feedback: when the robot makes a sound, the sensor will register the sound with some delay, creating spurious contingencies.
- Variability in background conditions: if the baby robot is in a noisy room, the sensor will be frequently active. If it is in a quiet room, the sensor will be seldom active. The baby robot needs to consider the level of background activity when deciding whether or not a social agent is present.
- Variability in social agents' responsiveness: social beings are highly unpredictable, with different individuals having different levels of responsiveness. The baby robot needs to consider the potential levels of activity of the agent when deciding whether or not an agent is present.

These considerations point to three causal factors that activate the baby robot's sensor: (1) self-feedback, (2) background activity independent of the robot, and (3) responsive social agents. The baby robot may find itself in one of two possible situations, or *contingency clusters*, which we identify with the following names: "responsive agent absent," and "responsive agent present" (see Fig. 3). When the robot makes a sound and no responsive agent is present, the robot's auditory sensor will activate for a period of time due to self-feedback. Afterward, the sound sensor becomes active at random times due to background activity. In addition to the self-feedback and background periods, there is a critical period of time during which social agents will respond to the robot's sounds, but only if a responsive agent is present (see Fig. 3).

3.1. State, action, observation, system dynamics, and sensor model:

The state X_t of the baby robot's environment contains five relevant variables: S, Z_t, K_1, K_2, K_3. The first variable S encodes whether a responsive agent is present ($S = 1$) or absent ($S = 0$). The second variable Z_t is a timer that encodes the amount of time since the baby robot's last vocalization. This timer determines which of three periods the baby robot is in: (1) a self-feedback period, which occurs immediately after a sound is made, (2) a critical period, during which social agents are more likely to respond to the last baby robot vocalization, and (3) a background period, unlikely to contain responses to the last vocalization. These three time periods are defined by the parameters $0 \leq \tau_1^s \leq \tau_2^s < \tau_1^a \leq \tau_2^a$. If the timer Z_t is between τ_1^s and τ_2^s, then the robot is in its self-feedback period. If the timer takes a value between τ_1^a and τ_2^a, then the system is in the critical period, during which social agents are likely to respond to the last vocalization. If Z_t is larger than τ_2^a, then the observed sounds are unlikely to be related to the last baby robot vocalization (See Fig. 3). The last three state variables K_1, K_2, K_3 are real-valued numbers that represent the expected rates of sensor activity during self, agent, and background periods. The state variables S, K_1, K_2, K_3 are assumed to be static. The timer variable Z_t increases by one on each time step until the baby robot vocalizes, at which point it resets to 1.

The action U_t represents the activation of the robot's sound actuator (e.g., a loudspeaker). At each moment, the baby robot can choose to vocalize ($U_t = 1$), meaning that it will activate the loudspeaker at time t. Otherwise it can choose to not vocalize, i.e., deactivate the loudspeaker ($U_t = 0$). By choosing the "vocalize" action, the baby robot is implicitly choosing to reset the timer Z_t that governs the unfolding of natural, social turn-taking behavior. By choosing the "do not vocalize" action, the baby robot is choosing to let the timer run its course.

We let Y_t represent the activation of the baby robot's sound sensor (e.g., a microphone). $Y_t = 1$ indicates that the sound level is larger than a fixed threshold θ, otherwise $Y_t = 0$. At each time step the sensor activates in a probabilistic manner. The probability that it becomes active is determined by K_1, K_2, K_3. If the timer Z_t is such that the system is in the self-feedback period, then the probability of activation is K_1. If Z_t is such that the system is in the critical period of agent response, then the probability of activation is K_2. Otherwise the system is in the background period and the probability of activation is K_3. If an agent is present and responding ($S = 1$) then K_2 and K_3 will be different. If an agent is not present ($S = 0$), then the agent and background activity rates are the same, i.e., $K_2 = K_3$.

Under this model, the problem of detecting that a responsive agent is in the room is equivalent to the problem of detecting whether the background time K_3 and agent time K_2 rates of sensory activation are different.

3.2. Inference process

The baby robot is assumed to follow an optimal probabilistic inference process. The specifics of this process are explained in the Appendix. For now, it suffices to say that at every point in time t, this process correctly determines the probability $p(s \mid h_t)$ that a social agent is responding s given the history of vocalizations and sounds h_t. If this probability is close to 0.5, the robot is uncertain about the presence or absence of a contingency. If $p(S = 1 \mid h_t) \approx 1$, the robot is quite certain that a responsive agent is present. If $p(S = 1 \mid h_t) \approx 0$, the robot is quite certain a responsive agent is not present. A common measure of the level of uncertainty about a random variable is the entropy, in bits, of the probability distribution of that variable, i.e.,

$$\mathcal{H}(S \mid h_t) = -\sum_{s=0}^{1} p(s \mid h_t) \log_2 p(s \mid h_t). \tag{5}$$

For example, if $p(S = 1 \mid h_t) = 0.5$ then the entropy is 1 bit (high uncertainty). If $p(S = 1 \mid h_t) = 0.99$ or $p(S = 1 \mid h_t) = 0.01$ then the entropy is 0.08 bits (low uncertainty).

3.3. Goal: information maximization

The goal of the baby robot is to gather as much information as possible and as quickly as possible about S, i.e., about the presence or absence of a social contingency. We call control policies that are optimized for the goal of information gathering "information maximization controllers" (infomax controllers for short).

Suppose by time t, the robot has access to the history h_t of sensor data and actions performed up to that time. A natural way to define an infomax controller is to let the reward at time t be equal to the amount of information that h_t provided about S, i.e.,

$$r_t = \mathcal{I}(S, h_t) \tag{6}$$

where $\mathcal{I}$ is the mutual information operator, an information theoretic quantity that corresponds to the intuitive notion of "information about" (see Appendix). Mutual information encodes the amount of information that the history of observed data h_t provides about the state S. This information can be expressed as a difference of entropies,

$$\mathcal{I}(S, h_t) = \mathcal{H}(S) - \mathcal{H}(S \mid h_t) \tag{7}$$

where $\mathcal{H}(S)$ is the initial uncertainty (entropy) about S, i.e., how uncertain the baby robot is about whether or not a social agent is present and responding *before it has done or heard anything*. This initial uncertainty is a constant independent of the available data and thus independent of the controller. $\mathcal{H}(S \mid h_t)$ is the uncertainty about S given the available data h_t. This value depends on the data history h_t, and therefore on the controller. Since $\mathcal{H}(S)$ is independent of the controller, we can ignore it and simply use the following reward function:

$$R_t = -\mathcal{H}(S \mid h_t). \tag{8}$$

This reward function promotes controllers that choose vocalizations that lead the baby robot to have high confidence (low entropy) about S. From a pure infomax standpoint, Baby-9 did not necessarily care that the social agent was responding to him. Instead, he cared about knowing whether or not it was responding to him. He would be just as happy after discovering that the unresponsive outcome was the correct one, just so long as he was confident in that discovery.

This brings us to the first challenge problem:

- What does it mean to "ask questions" for an organism like Baby-9 that does not have language?

From an infomax point of view, questions are behaviors that are expected to provide information about variables of interest. We hypothesize that Baby-9 was asking about the state variable S: "Is that thing out there responding to me?" To say that Baby-9 was asking questions about the state S means that his vocalizations helped him resolve his uncertainty about S. To say that he was asking *good* questions about S means that his vocalizations helped him resolve his uncertainty about S as quickly as possible. In order to analyze whether Baby-9 was doing something smart, i.e. asking good questions, we must first find the optimal controller, and then compare Baby-9's behavior with that of the optimal controller.

4. Optimal infomax controller for detecting social contingencies

4.1. Model parameters

The model described above has the following parameters:

- Δ_t: the sampling period used to discretize time.
- $\tau_1^s \leq \tau_2^s < \tau_1^a \leq \tau_2^a$: latency parameters that determine the self-feedback period, the period for agent likely responses, and the background period.

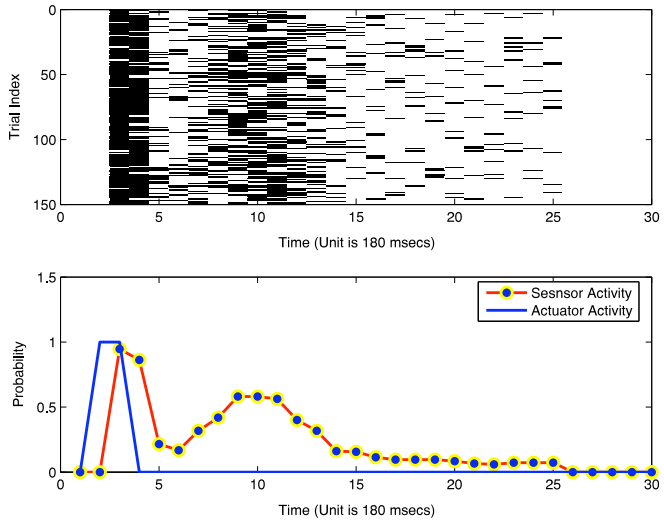

Fig. 4. Top: raster plot of 150 trials. On each trial a robot made a sound and subjects were asked to talk back to the character and let it know that they were listening. Dark indicates that the audio sensor was active. Bottom: probability of the audio sensor being active as a function of time. The probabilities are estimated by averaging across the 150 trials in the raster plot.

- θ: the threshold used to binarize the output of a sound sensor.
- π: the probability that an agent is present, prior to collecting any data.
- The time horizon τ over which the controller optimizes the information reward.

In order to set these parameters to reasonable values we conducted a study with four people that played the role of caregivers. They were presented with a humanoid robot that made sounds at randomly selected intervals. The participants were asked to treat the robot as if it were a baby, and to respond verbally to the sounds it made. The ages of the participants were 4, 6, 24, and 35 years. Each participant interacted independently with the robot for a five minute period. During this time, the robot vocalized at random intervals and the participants responded to it in the way that was most natural to them.

There were a total of 150 trials, during which the vocalizations of the robot and participants were digitized. Each trial started with a vocalization of the robot and ended 4 s later. The sound intensity threshold θ was chosen automatically by applying a k-means clustering procedure to the digitized sound data.

Fig. 4 shows the probability of activation of the binarized sound sensor as a function of time over 150 trials. The first peak in activity of the sound sensor is due to self-feedback, i.e., the sensor is recording its own sound. This peak occurs at 360 ms, indicating a delay between the time at which the program told the robot to make a sound, and the time at which the sound was detected by the sound sensor. By about 1300 ms after the end of the robot's vocalization, there is a second, smaller peak of activity in the sensor, which is now caused by the vocalizations of the human participants.

We chose Δt large enough to make self-feedback delays negligible, thus fixing $\tau_1^s, \tau_2^s = 0$, but small enough to capture the behaviors of interest. We found $\Delta t = 800$ ms to be a good compromise. The limits for the agent activation intervals were set to $\tau_1^a = 1$, $\tau_2^a = 3$, i.e., 800 ms and 2400 ms respectively. The prior probability that an agent is present was set to $\pi = 0.01$, thereby requiring a significant amount of data to become convinced that an agent is present. The reason for choosing this conservative value for π is explained below. The time horizon parameter τ was set to 40 time steps, i.e., 32 s. This value was chosen because, at the time, it was the longest horizon for which we could compute an optimal controller in a reasonable amount of time. As we explain later, we found that after approximately 12 time steps (approximately 10 s) the controllers stabilize. This indicates that it does not pay to use horizons longer than 10 s in situations governed by the statistics of social interaction.

4.2. Computation and analysis of the optimal controller

Infomax control is a specific instance of a general class of control problems known as partially observable Markov decision processes (POMDPs). In infomax control, information gain acts as a reward signal. The utility function optimized by the controller is the long term gathering of information about states of the world that are not directly observable. While finding exact solutions to infomax control problems is generally difficult, in this particular case there is a recursive statistic A_t that summarizes the observable data history without any loss of information. This allowed us to find an optimal controller using standard dynamic programming algorithms (Bertsekas, 2007) (See Appendix C).

The solution found using dynamic programming was a large lookup table that mapped each possible statistic a_t of the sensory–motor history h_t into a binary action u_t. Such a lookup table is provably optimal for every possible state, but it does not give us much intuition about which features of the sensory–motor history were important for making the optimal decision. In order to gain a better understanding of how the controller solved the problem, we developed a simple model that was evaluated on its ability to predict what the optimal controller would do next. We focused on the behavior of the controller for time steps $18 \leq t \leq 24$, because these are times that are not too close to the beginning and end of the controller's window of interest. We found that the following control policy matched the action of the optimal controller with 98.5% accuracy over all possible data history conditions:

$$c_t(h_t) = \begin{cases} 1 & \text{if } Z_t > \tau_2^a \text{ and } \dfrac{\mathrm{Var}(K_2 \mid h_t, S_t = 1)}{n_{2,t}} \\ & \quad > 9\dfrac{\mathrm{Var}(K_3 \mid h_t, S_t = 1)}{n_{3,t}} \\ 0 & \text{else} \end{cases} \tag{9}$$

where Z_t is the time since the last vocalization of the robot, and $\mathrm{Var}(K_2 \mid h_t, S_t = 1)$ is the current uncertainty (variance) about K_2, the sensor activation rate during the critical period in which social agents respond to the robot's vocalizations. $\mathrm{Var}(K_3 \mid h_t, S_t = 1)$ is the current uncertainty (variance) about K_3, the sensor activation rate during background noise periods, i.e. periods under which social agents are unlikely to respond to the last vocalization of the robot. The denominators dividing the variances indicate the total number of time steps collected up to date for the agent period ($n_{2,t}$) versus the background period ($n_{3,t}$). Dividing the variance by the number of observations accounts for how much the variance can be expected to reduce further with new observations.

Thus, the optimal controller always waits at least τ_2^a s, the longest period of time under which agents are likely to respond, before making a new vocalization. In addition, it does not vocalize unless it is significantly more uncertain about the rate of sensor activation during the critical period of social response than about the rate of activation during background periods. The effect is to homeostatically keep the uncertainty about the agent interval and the uncertainty about the background interval at a fixed ratio. If the agent rate is too uncertain, then the controller chooses to vocalize, thereby earning an opportunity to learn more about the rate of the agent intervals. If the background rate is too uncertain, then the controller chooses to remain silent, thereby gaining information about background intervals.

Notably, for a vocalization to occur, the uncertainty about the sensor activation rate K_3 during the agent period has to be at least 9 times larger than the uncertainty about the rate during the background period K_2. This may be due to the fact that vocalizations are more costly, in terms of information return, than silent periods. If the baby robot chooses to vocalize at time t, it gains no information during the times $[t + \tau_1^s, t + \tau_2^s]$ since self-feedback observations are not informative about S. In addition, during times $[t + \tau_1^a, t + \tau_2^a]$ the controller instructs the robot not to act and thus during those periods the robot can only gain information about K_2, not K_3. By contrast if the robot chooses to remain silent at time t, no time will be wasted due to self-feedback. Moreover, the robot can still choose to act or not to act in the future without constraints. This helps explain why uncertainty about the agent activity rate K_2 needs to be much larger than the uncertainty about the background activity rate, K_3, before an action occurs.

Note that "greedy" one-step controllers (Nelson & Movellan, 2001; Nelson et al., 2001) that seek as much information reward as possible immediately, at the expense of future expected rewards would fail on this task. The reason is that when the baby robot chooses to vocalize, its self-vocalization prohibits it from getting any information about K_2 or K_3 temporarily, while it would still get a small amount of information about K_3 by choosing to remain silent. Thus a greedy controller ends up deciding to never vocalize. Looking into the future allows the baby robot to conclude that vocalizing periodically provides a better long term information return than always choosing silence.

4.3. Comparison with the behavior of Baby-9

We compared the behavior of the optimal infomax controller described above to the behavior observed in the video of Baby-9. This video lasts 43 s, during which Baby-9 produced 7 vocalizations. The first vocalization occurred 5.58 s into the experiment. The intervals, in seconds, between the beginning of two consecutive infant vocalizations were as follows: {4.22, 10.32, 5.32, 6.14, 5.44, 3.56}. Most observers report that Baby-9 clearly has detected that there is a responsive agent in the room by the end of the 43 s.

We ran the optimal controller with a receding time horizon of 24 time steps (19.2 s), i.e., at each point in time the controller behaved so as to maximize the expected information to be gained over a period of 19.2 s into the future. As in the Baby-9 experiment, every time the baby robot's controller made a sound, it was given a response, simulating a social agent. Fig. 5 shows the result of the simulation.

The top graph shows the vocalizations of the optimal controller, which serves as a model of Baby-9. The infomax controller exhibited turn-taking behaviors that were very similar to the ones observed in Baby-9: the infomax controller makes a sound and follows it by a period of silence as if waiting for the outcome of a question.

This turn-taking behavior was not built into the system. Instead, it emerged from the requirement to maximize information gain given the time delays and levels of uncertainty typical in social interactions.

The controller produced six vocalizations over a period of 43 s. The average interval between vocalizations was 5.92 s which is remarkably close to the average of 5.83 s of silence between vocalizations for Baby-9. There seems to be a tendency both in the model and in Baby-9 for the early silence intervals to be longer than the later ones.

This provides an answer to the second challenge problem in the introduction to this document:

- Was it smart for Baby-9 to schedule his vocalizations in the way that he did?

Baby-9's behavior was smart in the sense that he asked good questions: questions that helped to quickly resolve his uncertainty about whether the rectangular prism in front of him was actually a contingent social agent. In fact, Baby-9's pattern of vocalizations and silences was very close to optimal. This also explains the sense of intentionality that most people intuitively perceive when they watch the video of Baby-9. The behavior of Baby-9 makes a great deal of sense if one were to assume that his goal is to discover whether social agents are responsive to him.

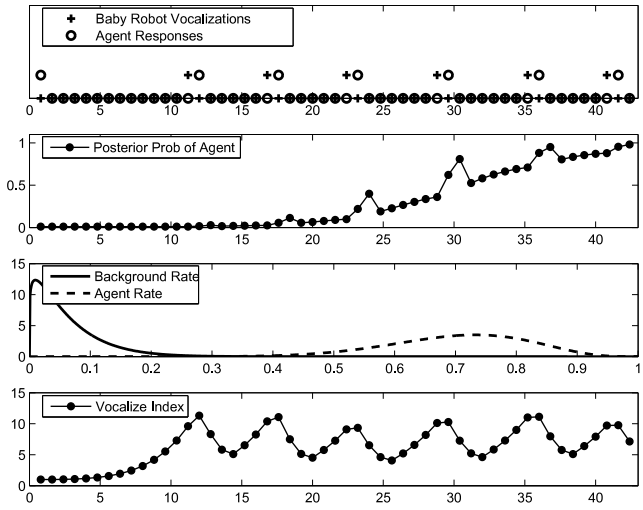

Fig. 5. The horizontal axis represents time in seconds. From top to bottom: (1) Responses of the infomax controller (which simulates a baby). Note that the social agent responded every time the baby robot vocalized, but otherwise the environment was silent. (2) Posterior probability for the presence of a responsive agent as a function of time. (3) Posterior distribution for the agent and background rates after 43 s. (4) Ratio of the uncertainty about the agent's response rate versus the uncertainty about the background's response rate.

The second graph from the top in Fig. 5 shows the system's beliefs about the presence of a responsive agent. These beliefs are updated in real time using standard Bayesian inference (see Appendix). In our simulation, we chose a conservative prior probability $\pi = 0.01$ for the presence of social contingency to force the controller to gather a significant amount of data before deciding that there is a social contingency present. Note that in spite of this conservative prior, by the end of the 43 s, the posterior probability that there is a responsive agent is very close to 1. The third graph shows the posterior probability distributions about the agent and background response rates by the end of the 43 s period. Note that these two distributions are very different, consistent with the idea that there is indeed a responsive agent present.

This provides an answer to the third challenge problem in the introduction:

- Was it smart for him to decide within a few responses and less than a minute into the experiment that the robot was responsive?

Given the statistics of social interaction, it was indeed very smart for Baby-9 to decide within a few responses and less than a minute into the experiment that a social contingency was present.

Finally, the last graph in Fig. 5 shows the ratio between the uncertainty about K_2, the sensor rate during agent periods, and the uncertainty about K_3, the sensor rate during background periods. Note that when this ratio reaches the value of 9, the optimal controller vocalizes.

5. Learning to detect contingencies

In the previous section, we used standard dynamic programming algorithms to find an optimal infomax controller. We found that this model appeared to describe well the turn-taking behaviors observed in some 10-month-old infants when they are trying

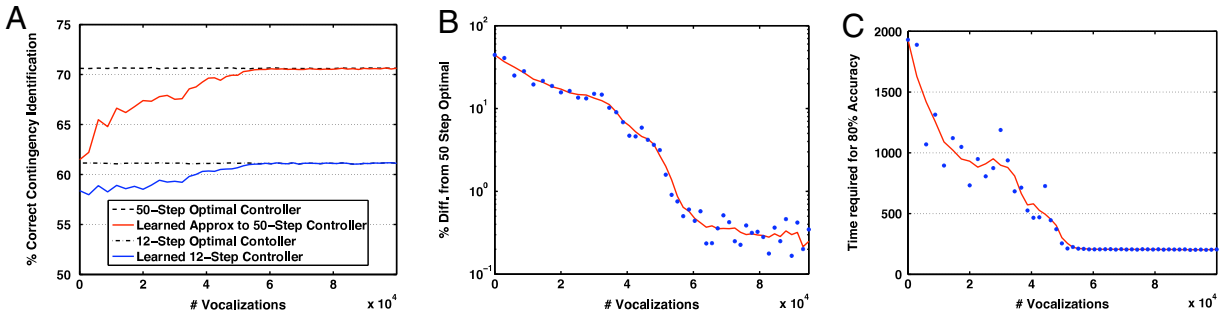

Fig. 6. (A) Performance of infomax TD Learning in the finite horizon (12-step), and receding-horizon (50-step) case, based on the total number of vocalizations made since birth. (B) When a receding-horizon controller with 6.5 s of memory and a 3.5 s deadline is used to approximate an optimal controller with a perfect memory and much longer deadline, the final information gathering performance is nearly identical. (C) The number of time steps spent acting, exploring, and listening to the world that are required to achieve 80% social agent identification accuracy.

to detect the presence of social contingency. This begs the question: how did these infants acquire a policy for finding social contingency that is so close to optimal?

One possibility is that children are born with these policies. The differences in contingency detection efficiency found between 2-month-old infants and 10-month-old infants may be due to the maturation of brain structures. Just like teeth mature to allow more efficient chewing, some brain structures may be specially programmed by evolution to mechanistically mature into a machine for more efficient detection of contingency.

Another possibility is that children are born with something akin to a dynamic programming algorithm that allows them to find the optimal controller. The advantage of dynamic programming is that it finds controllers guaranteed to be optimal. However, the dynamic programming hypothesis has several drawbacks: (1) it requires detailed and precise knowledge of the system dynamics and observation model; (2) it is very time and memory intensive; (3) it is not easily implementable on neural-like hardware; (4) it provides no mechanisms to benefit from experience interacting with the world.

An alternative to both pre-programmed controllers and dynamic programming is reinforcement learning (RL). RL is an area of machine learning and control in which the goal is to learn control policies that approximate the solutions given by dynamic programming without requiring detailed and precise knowledge of the system dynamics (Sutton & Barto, 1988). RL is easily implementable in neural-like hardware and provides a natural set of mechanisms to make good use of experience and interaction in the world. Unfortunately, RL itself has drawbacks. While dynamic programming requires a good deal of time, memory, and computational effort, RL requires many trial and error experiences to learn efficient policies. In a sense, the difficulty of the computation is offloaded to the world around the robot, and to interaction with its environment. The amount of experience required in some cases is so great that RL cannot be considered as a plausible model of learning in a developmentally reasonable time frame.

5.1. Infomax RL results

In this section, we consider whether the optimal contingency detection strategies observed in some 10-month-old infants could be explained as the manifestation of an RL process driven by an information based reward system (infomax RL). To demonstrate the computational plausibility of the infomax RL hypothesis, it suffices to show that at least one RL algorithm can learn within a developmentally plausible period of time. We chose this time frame to be 60,000 vocalizations, which was meant to be a conservative ballpark estimate, based on 200 vocalizations per day of the infant's first 10 months of life.

We implemented infomax RL using temporal difference (TD) learning, a popular RL algorithm that has been shown to have correspondences in the pattern of dopamine release from neurons in the basal ganglia (Schultz, Dayan, & Montague, 1997) (See Appendix D).

Empirically, we found that the number of vocalizations needed for the TD learning algorithm to converge grew as a fifth power of the temporal horizon τ. Convergence within 60,000 vocalizations was only achievable with horizons of 12 time steps (10 s) or less. A horizon of 16 time steps required 230,000 vocalizations, and a horizon of 20 time steps required 700,000 vocalizations, which is much higher than our estimate of a reasonable developmental time frame.

We then investigated the question of how 10 s controllers compare optimal controllers with longer time horizons. Given the statistics of social interaction, does it pay off to use time horizons longer than 10 s?

Fifty new simulations were performed, each with different starting points and with a time horizon of 12 time steps. As expected, on average, infomax RL converged after less than 60,000 vocalizations. We then used dynamic programming to compute optimal 12-step and 50-step controllers in order to serve as evaluation standards for the controller learned from experience. The performance of the optimal 12-step controllers found using dynamic programming (an exact method) was identical to the 12-step controllers found using infomax RL (an approximate method), indicating that infomax RL converged to an optimal solution.

To compare the learned controller to the 50-step optimal controller, we adopted a receding-horizon approach: the 12-step learned controller was artificially limited to eight time steps of memory (about 6.5 s), and then chose the action that would help it gather as much information as possible in the next four time steps (about 3.5 s). This limited memory controller, which had learned from experience and information reward over a simulated ten month time frame, was almost as good as the performance of the optimal 50-step controller: after 60,000 vocalizations, the average performance was better than 99.5% of the optimal performance (see Fig. 6(A) and (B)).

This provides an answer to our fourth and final challenge problem:

- What mechanisms can explain the transition from the relatively slow learning that Watson observed in 2-month-old infants to the very fast and active learning that was observed in 10-month-old infants like Baby-9?

Simple reinforcement learning algorithms, in which uncertainty reduction is used as a reward signal, are a plausible mechanism to explain how infants improve on their capacity to detect contingencies. In 10 months of simulated experience, infomax RL agents perform 99.5% as well as the best possible controller.

6. Real-time robot implementation

Once computed, the optimal infomax policy can be applied to sensor data in real time, trivially, on any modern computer. To test how well this policy would work in real life, we implemented it on RobovieM, a humanoid robot developed at ATR's Intelligent Robotics and Communication Laboratories. While the robot was not strictly necessary to test the real-time controller, it greatly helped to improve the quality of the interactions developed between humans and machines, thereby providing a more realistic method for testing the controller.

For the binary sensor, we chose to average acoustic energy over 500 ms windows and binarize it using the threshold θ that was found by applying a k-means algorithm to the acoustic portion of the natural interaction data that were collected previously. The actuator was a small loudspeaker producing a 200 ms robotic sound. The self-feedback delay parameters of the controller were chosen by measuring the time delay between issuing a command to produce a sound and receiving feedback from the audio sensor. The agent delay parameters were the same as in the simulation of Baby-9.

The robot was programmed to change its posture based on the controller's belief about the presence/absence of a responsive agent: a posture that indicated a high level of attention when the controller believed that an agent was present, and a posture that indicated boredom when it believed that an agent was not present.

Overall, the infomax controller was remarkably effective in a wide range of environments, and it required very little computational and sensory resources. In standard office environments, with relatively high levels of noise, the controller reliably detects within 3 or 4 vocalizations whether or not a responsive agent is present. We have demonstrated this system at both scientific talks and poster sessions. Demonstrations at talks, which generally have relatively low noise levels, work very well. During poster sessions, the rooms are typically very noisy, but it only takes a few more vocalizations for the controller to gather enough information to make reliable decisions. The level of performance is remarkable considering the difficulty of these adverse conditions, and the simplicity of the sensors being used.

7. Conclusions

There is evidence that the ability to detect social contingencies plays an important role in the social and emotional development of infants (Bigelow, 1999; Bahrick & Watson, 1985; Watson, 1972, 1979, 1985). Analyzing this problem from a computational perspective provided important clues for understanding social development in infants and for the synthesis of social behavior in robots. We framed our analysis of contingency detection within the theory of stochastic optimal control. In particular, we formulated contingency detection as a control problem in which the goal is to gather information as efficiently as possible about the presence or absence of contingencies.

A popular model of the social contingency detection problem describes social agents and background noise as Poisson processes (Watson, 1985). We showed that under this model, the optimal information gathering policy exhibits turn-taking behaviors very similar to the ones found in some 10-month-old infants: vocalizations followed by periods of silence of about 6 s. The results suggest that some 10-month-old infants have an exquisite understanding of the statistics of social interaction and have acquired efficient policies to operate in this world. Even though these infants lack a language, they are already asking questions: they schedule their vocalizations in a manner that maximizes the expected information return given the temporal statistics of social interaction.

One of our goals was to explore to what extent social development can be bootstrapped from simple perceptual and learning primitives so that it can be synthesized in robots. For example, our approach does not require high-level conceptual primitives, such as the concept of people or the idea that people have minds. In our model, the terms "responsive agent present" and "responsive agent absent" are just mnemonic labels for contingency clusters that may not correspond to categories easily describable with words. Indeed, in John Watson's original experiment (Watson, 1972), 2-month-old infants seemed to group together responsive caregivers and contingent mobiles.

We showed that simple temporal difference reinforcement learning mechanisms could explain how infants acquire the efficient social contingency detection strategies observed in some 10-month olds. The key is to use the reduction of uncertainty (information gain) as a reward signal. The result is an interesting form of learning in which the learner rewards itself for conducting actions that help reduce its own sense of uncertainty. Traditional models of classical and operant learning emphasize the role of external reward stimuli, like food or water. The brain is probably set up to recognize these stimuli and to encode them as rewarding because it is advantageous to do so. Infomax control suggests that it may also be similarly advantageous for organisms to recognize uncertainty and to encode the reduction of uncertainty as rewarding. There is some evidence that the brain may indeed reward reduction in uncertainty with the same mechanisms that it rewards food or water. It has been found that dopamine-releasing neurons located in the substantia nigra pars compacta and ventral tegmental area play an important role in reward based learning (Montague, Hyman, & Cohen, 2004; Schultz et al., 1997; Wise, 2004). Initially the activity of these neurons was studied for basic forms of reward, such as food and water. However, in recent years it has been found that the same neurons that signal the expected amount of physical rewards, like food or water, also signal expected information gain. Thus it appears that information gain may indeed have a special status as an intrinsic motivational reward in the brain (Bromberg-Martin & Hikosaka, 2009).

The long term goal of this paper is to illustrate the possibilities of a science of behavior and development that is anchored on rigorous computational analysis. As proposed by Edelman and Vaina (2001) and Marr (1982), the goal of computational approaches is to help understand the problems faced by the brain, as well as the solutions it finds, when operating in everyday life. This approach offers a modern alternative to the behaviorist and the mentalist/cognitive approaches that dominated psychology in the 20th century.

Computational analysis has proven to be a very useful tool for the study of the brain. Our hope is to illustrate that it may also prove useful to understand social development, and to synthesize it in robots. It is remarkable that, after all these years, neither the behaviorist nor the cognitive/mentalist traditions in psychology have significantly contributed to the synthesis of intelligent behavior. We believe that stochastic optimal control may provide a formal mathematical foundation for an emerging area of computer science and engineering that focuses on the computational understanding of human behavior, and on its synthesis in robots.

Acknowledgements

The researchers were sponsored by the NSF IIS INT2-Large and the NSF Science of Learning Center grant SBE-0542013.

Appendix A. Definitions and conventions

Unless otherwise stated, capital letters are used for random variables, small letters for specific values taken by random

variables, and Greek letters for fixed parameters. When the context makes it clear, we identify probability functions by their arguments: e.g., $p(x, y)$ is shorthand for the joint probability mass or joint probability density that the random variable X takes the specific value x and the random variable Y takes the value y. We use subscripted colons to indicate sequences: e.g., $X_{1:t} \stackrel{\text{def}}{=} \{X_1 \cdots X_t\}$. We work with discrete time stochastic processes, with the parameter $\Delta t \in \mathbb{R}$ representing the sampling period. We use E for expected values and Var for variance. The symbol $\sim$ indicates the distribution of random variables. For example, $X \sim \text{Poisson}(\lambda)$ indicates that X has a Poisson distribution with parameter λ. We use $\delta(\cdot, \cdot)$ for the Kronecker delta function, which takes value 1 if its two arguments are equal, otherwise it takes value 0.

- Beta Variables:

$$X \sim \text{Beta}(\beta_1, \beta_2) \tag{10}$$

$$\begin{aligned} p(x) &= \text{Beta}(x, \beta_1, \beta_2) \\ &= \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)}(x)^{\beta_1 - 1}(1 - x)^{\beta_2 - 1} \end{aligned} \tag{11}$$

$$\text{E}(X) = \frac{\beta_1}{\beta_1 + \beta_2} \tag{12}$$

$$\text{Var}(X) = \frac{\beta_1 \beta_2}{(\beta_1 + \beta_2)^2(\beta_1 + \beta_2 + 1)} \tag{13}$$

where Γ is the Gamma function

$$\Gamma(z) = \int_0^\infty t^{z-1} \text{e}^{-t} \text{d}t. \tag{14}$$

- Entropy:

$$\mathcal{H}(Y) = -\int p(y) \log p(y) \text{d}y. \tag{15}$$

- Conditional Entropy:

$$\mathcal{H}(Y \mid x) = -\int p(y|x) \log p(y \mid x) \text{d}y \tag{16}$$

$$\mathcal{H}(Y \mid X) = -\int p(x, y) \log p(y \mid x) \text{d}x\text{d}y \tag{17}$$

$$= \int p(x) \mathcal{H}(Y \mid x) \text{d}x. \tag{18}$$

- Mutual Information: the information about the random variable Y provided by the specific value x from the random variable X is defined as follows:

$$\mathcal{I}(Y, x) = \mathcal{H}(Y) - \mathcal{H}(Y \mid x). \tag{19}$$

The average information about the random variable Y provided by the random variable X is defined as follows

$$\mathcal{I}(Y, X) = \sum_x p(x) \mathcal{I}(Y, x) = \mathcal{H}(Y) - \mathcal{H}(Y \mid X). \tag{20}$$

Appendix B. Summary of the contingency detection model

Parameters: The extended version of the model has 15 parameters:

$\Delta t \in \mathbb{R}$. Sampling period in seconds.

$\pi \in [0, 1]$. Prior probability.

$0 \leq \tau_1^s \leq \tau_2^s$. Delay parameters for self-feedback loop.

$\tau_2^s < \tau_1^a \leq \tau_2^a$. Delay parameters for social agents.

$(\beta_{i,1}, \beta_{i,2}), i = 1, 2, 3$. Parameters for Beta Prior distribution.

θ. Threshold for binarizing auditory signal. (21)

τ. Time horizon.

For the simulations presented in this paper, we worked with a simplified model with 5 parameters: $\Delta t, \tau_1^a, \tau_2^a, \theta, \pi$. We choose Δt large enough to make delays in the onset of self-feedback to be negligible, thus fixing $\tau_1^s, \tau_2^s = 0$, but small enough to capture the behaviors of interest. We found $\Delta t = 800$ ms to be a good compromise. The values of τ_1^a were set based on a pilot study described in the main part of this paper: $\tau_1^a = 1, \tau_2^a = 3$, i.e., 800 ms and 2400 ms respectively. In the simplified model, we treat the agent and background response rates as random variables with uninformative priors, thus fixing the β parameters to 1. We chose $\pi = 0.01$ thus making the prior probability for the presence of agents small, requiring large likelihood ratios to become convinced that an agent is present. The sound threshold θ was chosen using a k-means maximum entropy procedure on the statistics of the available sound. We chose the largest temporal horizon $\tau = 40$ for which we could compute an optimal controller using traditional dynamic programming approaches. Later investigation showed that longer time horizons do not significantly change the optimal policy.

Static random variables:

$S \sim \text{Bernoulli}(\pi)$. Presence/Absence of Responsive Agent (22)

$K_1 \sim \text{Beta}(\beta_{1,1}, \beta_{1,2})$.
Sensor activity rate during self period. (23)

$K_2 \sim \text{Beta}(\beta_{2,1}, \beta_{2,2})$.
Sensor activity rate during agent period (24)

K_3. Sensor activity Rate during background period (25)

$K_3 \sim \text{Beta}(\beta_{3,1}, \beta_{3,2})$, if $S = 1$ (26)

$K_3 = K_2$, if $S = 0$. (27)

Stochastic processes:

The following processes are defined for $t = 1, 2, \ldots$

$$\text{Timer:} \quad Z_t \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } U_{t-1} = 1 \\ Z_{t-1} + 1 & \text{if } U_{t-1} = 0 \text{ and } Z_{t-1} \leq \tau_2^a \\ Z_t & \text{else.} \end{cases}$$

$$\text{Indicator of Self Period:} \quad I_{1,t} = \begin{cases} 1 & \text{if } Z_t \in [\tau_1^s, \tau_2^s] \\ 0 & \text{else.} \end{cases}$$

$$\text{Indicator of Agent Period:} \quad I_{2,t} = \begin{cases} 1 & \text{if } Z_t \in [\tau_1^a, \tau_2^a] \\ 0 & \text{else.} \end{cases}$$

Indicator of Background Period: $I_{3,t} = (1 - I_{1,t})(1 - I_{2,t})$.

Self Driver: $D_{1,t} \sim \text{Poison}(K_1)$.

Agent Driver: $D_{2,t} \sim \text{Poison}(K_2)$.

Background Driver: $D_{3,t} \sim \text{Poison}(K_3)$.

Robot Sensor: $Y_t = I_t \cdot D_t$.

Robot Controller: $C = (C_1, \ldots, C_\tau)$.

Robot Actuator: $U_t = C_t(Y_{1:t}, U_{1:t-1})$.

$$\text{Sensor Activity Counters:} \quad P_{i,t} = \sum_{s=1}^{t} I_{i,s} Y_s \quad \text{for } i = 1, 2, 3.$$

$$\text{Sensor Inactivity Counters:} \quad Q_{i,t} = \sum_{s=1}^{t} I_{i,s}(1 - Y_s)$$

for $i = 1, 2, 3$.

Appendix C. Detailed model description

The model presented in this section was inspired by Watson (1985) formulation of the social contingency detection problem: background and responsive caregivers are modeled as Poisson processes. Caregivers respond within a fixed window of time from

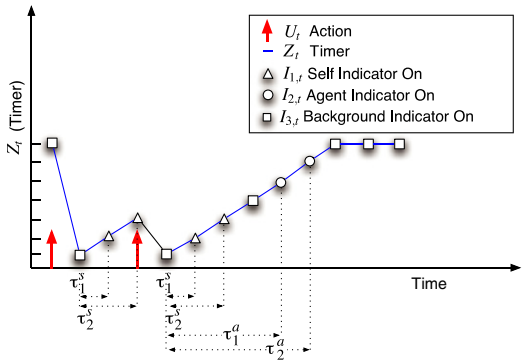

Fig. 7. Graphical representation of the dynamics of the timer and the indicator variables.

the last response from the baby. Watson focused on the inference problem, i.e., how to make decisions given the available data. Here we focus on the control problem, how to schedule behaviors in real time to optimally gather data.

C.1. Self-feedback processes

We let the robot sensor respond to its own actuators, e.g., the robot can hear its own vocalizations, and allow for delays and uncertainty in this self-feedback loop. In particular, we let the distribution of self-feedback delays be uniform with parameters $\tau_1^s \leq \tau_2^s$. The indicator variable for self-feedback period is thus defined as follows:

$$I_{1,t} = \begin{cases} 1 & \text{if } Z_t \in [\tau_1^s, \tau_2^s] \\ 0 & \text{else.} \end{cases} \tag{28}$$

During Self periods, the activation of the sensor is driven by the discrete time Poisson process $\{D_{1,t}\}$ that has rate K_1, i.e.,

$$p(D_{1,t} = 1) = K_1. \tag{29}$$

C.2. Social agent process

The parameters $0 \leq \tau_1^a \leq \tau_2^a$ bound the reaction times of social agents i.e., it takes agents anything from τ_1^a to τ_2^a time steps to respond to an action from the robot. "*Agent periods*", which are designated by the indicator process $\{I_{2,t}\}$ are periods of time for which responses of agents to previous robot actions are likely if an agent were to be present. The indicator variable for an agent period is as follows (see Fig. 7)

$$I_{2,t} = \begin{cases} 1 & \text{if } Z_t \in [\tau_1^a, \tau_2^a] \\ 0 & \text{else.} \end{cases} \tag{30}$$

During agent periods, the robot's sensor is driven by the Poisson process $\{D_{2,t}\}$ which has rate K_2, i.e.,

$$p(D_{2,t} = 1) = K_2. \tag{31}$$

The distribution of K_2 depends on whether or not a responsive agent is present. If an agent is present, i.e. $S = 1$, we let K_2 be independent of K_1 and K_3 and endow it with a prior Beta distribution with parameters $\beta_{2,1}, \beta_{2,2}$ reflecting the variability in response rates typical of social agents. If an agent is not present, i.e., $S = 0$, then the response rate during agent periods is the same as the response rate during background periods, i.e., $K_2 = K_3$.

C.3. Background process

The background is modeled as a Poisson process $\{D_{3,t}\}$ with rate K_3, i.e.,

$$p(D_{3,t} = 1) = K_3. \tag{32}$$

The background drives the sensor's activity that is not due to self-feedback and is not due to social agent responses. Note that this can include, among other things, the actions from external social agents who are not responding to the robot (e.g., two social agents may be talking to each other thus activating the robot's sound sensor). We let the background rate K_3 have a prior Beta distribution with parameters $\beta_{3,1}, \beta_{3,2}$ reflecting the variability of background activity from situation to situation. If $\beta_{3,1} = \beta_{3,2} = 1$ the distribution is uninformative, i.e., all responsiveness rates are equally possible a priori:

$$K_3 \sim \text{Beta}(\beta_{3,1}, \beta_{3,2}). \tag{33}$$

The background indicator keeps track of periods for which self-feedback or responsive actions from a social agent may not happen, i.e.,

$$I_{3,t} = (1 - I_{1,t})(1 - I_{2,t}). \tag{34}$$

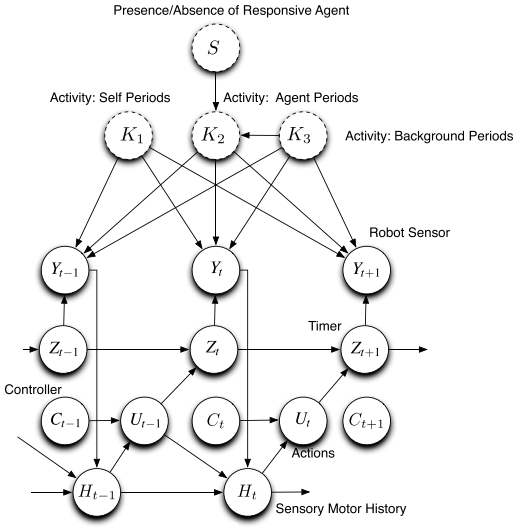

Fig. 8. Graphical representation of the model. Arrows represent dependency relationships between variables. Dotted figures indicate unobservable variables, continuous figures indicate observable variables. The controller C_t maps all the observed information up to time t into the action U_t. The effect of the action depends on the presence or absence of a responsive agent S and on the timing of the action as determined by Z_t. The goal is to maximize the information return about the actual value of S.

C.4. Sensor model

The activity of the sensor is a switched Poisson process: during self-feedback periods it is driven by the Poisson process $\{D_{1,t}\}$, during agent periods it is driven by $\{D_{2,t}\}$ and during background periods it is driven by $\{D_{3,t}\}$, i.e.,

$$Y_t = I_t \cdot D_t = \sum_{i=1}^{3} I_{i,t} D_{i,t}. \tag{35}$$

C.5. Auxiliary processes

We will use the processes $\{P_t, Q_t\}$ to register the sensor activity, and lack thereof, up to time t during self, agent and background periods. In particular for $t = 1, 2, \ldots,$

$$P_{i,t} = \sum_{s=1}^{t} I_{i,s} Y_s, \quad \text{for } i = 1, 2, 3 \tag{36}$$

$$Q_{i,t} = \sum_{s=1}^{t} I_{i,s} (1 - Y_s) \quad \text{for } i = 1, 2, 3. \tag{37}$$

C.6. Constraints

Fig. 8 displays the Markovian constraints in the joint distribution of the different variables involved in the model. An arrow from

variable X to variable Y indicates that X is a "parent" of Y. The probability of a random variable is conditionally independent of all the other variables given the parent variables. Dotted figures indicate unobservable variables, continuous figures indicate observable variables.

C.7. Optimal inference

Let $(y_{1:t}, u_{1:t}, p_t, q_t, z_t)$ be an arbitrary sample from $(Y_{1:t}, U_{1:t}, P_t, Q_t, Z_t)$. Then

$$p(y_{1:t} \mid k, u_{1:t}, s) = \prod_{i=1}^{3} (k_i)^{p_{i,t}} (1-k_i)^{q_{i,t}}. \tag{38}$$

Note that the rate variables K_1, K_2, K_3 are independent under the prior distribution. Moreover, if $S = 1$, then they affect the sensor at non-intersecting periods of time. It follows that the rate variables are also independent under the posterior distribution. In particular,

$$p(k \mid y_{1:t}, u_{1:t}, S=1) = \prod_{i=1}^{3} \text{Beta}(k_i; \beta_{i,1} + p_{i,t}, \beta_{i,2} + q_{i,t}). \tag{39}$$

If the null hypothesis is correct, i.e., $S = 0$, then $K_2 = K_3$, i.e., the probability distribution of sensor activity during the "agent periods" is the same as during background periods. Moreover, the set of times for which the sensor's activity depends on K_2, K_3 does not intersect with the set of times for which it depends on K_1. Thus K_1 will be independent of K_2, K_3 under the posterior distribution:

$$\begin{aligned} p(k \mid y_{1:t}, u_{1:t}, S=0) = {} & \text{Beta}(k_1; \beta_{1,1} + p_{1,t}, \beta_{1,2} + q_{1,t}) \\ & \text{Beta}(k_2; \beta_{2,1} + p_{2,t} + p_{3,t}, \beta_{2,2} + q_{2,t} + q_{3,t}) \delta(k_2, k_3). \end{aligned} \tag{40}$$

Note for an arbitrary k such that $p(k \mid y_{:t}, u_{1:t}, s) > 0$ we have that

$$\begin{aligned} p(y_{1:t} \mid u_{1:t}, s) &= p(y_{1:t} \mid k, u_{1:t}, s) \frac{p(k \mid u_{1:t}, s)}{p(k \mid y_{1:t}, u_{1:t}, s)} \\ &= p(y_{1:t} \mid k, u_{1:t}, s) \frac{p(k)}{p(k \mid y_{1:t}, u_{1:t}, s)}. \end{aligned} \tag{41}$$

Thus

$$\begin{aligned} & p(y_{1:t} \mid u_{1:t}, S=1) \\ &= \prod_{i=1}^{3} \left((k_i)^{p_{i,t}} (1-k_i)^{q_{i,t}} \frac{\text{Beta}(k_i; \beta_{i,1}, \beta_{i,2})}{\text{Beta}(k_i; \beta_{i,1} + p_{1,t}, \beta_{i,2} + q_{i,t})} \right) \end{aligned} \tag{42}$$

$$= \prod_{i=1}^{3} \frac{\Gamma(\beta_{i,1} + \beta_{i,2})}{\Gamma(\beta_{i,1})\Gamma(\beta_{i,2})} \frac{\Gamma(\beta_{i,1} + p_{i,t})\Gamma(\beta_{i,2} + q_{i,t})}{\Gamma(\beta_{i,1} + \beta_{i,2} + p_{i,t} + q_{i,t})} \tag{43}$$

and

$$\begin{aligned} p(y_{1:t} \mid u_{1:t}, S=0) = {} & \frac{\text{Beta}(k_1; \beta_{1,1}, \beta_{1,2})}{\text{Beta}(k_1; \beta_{1,1} + p_{1,t}, \beta_{1,2} + q_{1,t})} \\ & \times \frac{\text{Beta}(k_3; \beta_{3,1}, \beta_{3,2})}{\text{Beta}(k_3; \beta_{3,1} + p_{2,t} + p_{3,t}, \beta_{3,2} + q_{2,t} + q_{3,t})} \\ & \times \prod_{i=1}^{3} (k_i)^{p_{i,t}} (1-k_i)^{q_{i,t}} \end{aligned} \tag{44}$$

$$\begin{aligned} = {} & \frac{\Gamma(\beta_{1,1} + \beta_{1,2})}{\Gamma(\beta_{1,1})\Gamma(\beta_{1,2})} \frac{\Gamma(\beta_{1,1} + p_{1,t})\Gamma(\beta_{1,2} + q_{1,t})}{\Gamma(\beta_{1,1} + \beta_{1,2} + p_{1,t} + q_{1,t})} \\ & \times \frac{\Gamma(\beta_{3,1} + \beta_{3,2})}{\Gamma(\beta_{3,1})\Gamma(\beta_{3,2})} \\ & \times \frac{\Gamma(\beta_{3,1} + p_{2,t} + p_{3,t})\Gamma(\beta_{3,2} + q_{2,t} + q_{3,t})}{\Gamma(\beta_{3,1} + \beta_{3,2} + p_{2,t} + p_{3,t} + q_{2,t} + q_{3,t})} \end{aligned} \tag{45}$$

where we used the fact that $k_2 = k_3$ with probability one under $S = 0$. Thus the likelihood ratio between the two hypotheses is as follows:

$$\begin{aligned} L_t(p_t, q_t) &= \frac{p(y_{1:t} \mid u_{1:t}, S=1)}{p(y_{1:t} \mid u_{1:t}, S=0)} \\ &= \frac{\Gamma(\beta_{2,1} + \beta_{2,2})}{\Gamma(\beta_{2,1})\Gamma(\beta_{2,2})} \frac{\Gamma(\beta_{2,1} + p_{2,t})\Gamma(\beta_{2,2} + q_{2,t})}{\Gamma(\beta_{2,1} + \beta_{2,2} + p_{2,t} + q_{2,t})} \\ &\quad \times \frac{\Gamma(\beta_{3,1} + p_{3,t})\Gamma(\beta_{3,2} + q_{3,t})}{\Gamma(\beta_{3,1} + \beta_{3,2} + p_{3,t} + q_{3,t})} \\ &\quad \times \frac{\Gamma(\beta_{3,1} + \beta_{3,2} + p_{2,t} + p_{3,t} + q_{2,t} + q_{3,t})}{\Gamma(\beta_{3,1} + p_{2,t} + p_{3,t})\Gamma(\beta_{3,2} + q_{2,t} + q_{3,t})}. \end{aligned} \tag{46}$$

The posterior odds, which is the product of the prior odds and the likelihood ratio,

$$\frac{p(S=1 \mid y_{1:t}, u_{1:t})}{p(S=0 \mid y_{1:t}, u_{1:t})} = L(p_t, q_t) \frac{\pi}{1-\pi} \tag{47}$$

contains all the information available to the robot about the presence of a responsive agent.

C.8. Infomax control

The goal in infomax control is to find controllers that provide as much information as possible about a random variable of interest S. Suppose we have a fixed controller c under which we have observed the history of sensory–motor data $h_t = (u_{1:t-1}, y_{1:t})$. The information about the random variable S provided by the observed sequence is as follows:

$$\mathcal{I}(S, h_t) = \mathcal{H}(S) - \mathcal{H}(S \mid h_t). \tag{48}$$

The prior uncertainty $\mathcal{H}(S)$ does not depend on the observations, and thus it will be the same regardless of the controller c. Thus, if our goal is to gain information about S, then we can use as reward function the negative of the entropy of S given the observed sequence h_t, i.e.,

$$r_t \stackrel{\text{def}}{=} \mathcal{H}(S \mid h_t). \tag{49}$$

The value of a controller is expressed as a weighted sum of the expected accumulation of future rewards, up to a terminal time τ:

$$\rho(c) = \sum_{t=1}^{\tau} \alpha_t E[R_t \mid c] = \sum_{t=1}^{\tau} \alpha_t \mathcal{H}(S \mid Y_{1:t}, U_{1:t-1}) \tag{50}$$

where the $\alpha_t \geq 0$ are fixed numbers representing the relative value of information return at different points in time.

The controller c_t maps the information history $h_t = (y_{1:t-1}, u_{1:t-1})$ that is available prior to taking the action into the action taken at that time, i.e.,

$$u_t = c_t(h_t). \tag{51}$$

The information history is Markovian and the reward is a function of the information history. Therefore, infomax control is a Markov Decision process with respect to the information history. Unfortunately, the number of possible observable sequences grows exponentially as a function of time, making it very difficult to use standard optimal control algorithms for horizons beyond a few time steps. In particular each action and each observation is binary, i.e., for any given time t there are 2^{2t} separate state histories that must be learned. Fortunately the observation history can be summarized by a statistic A_t consisting of integers: the number of time steps since the last vocalization, the number of active and the number of inactive observations during the periods of agent and background states, i.e.,

$$A_t \stackrel{\text{def}}{=} (Z_t, P_{2,t}, P_{3,t}, Q_{2,t}, Q_{3,t}). \tag{52}$$

The statistic A_t has the following properties

1. It is a recursive function
$$A_{t+1} = f_t(A_t, U_t, Y_{t+1}). \tag{53}$$
2. The predictive distribution of Y_{t+1} is conditionally independent of H_t given A_t, U_t, i.e.,
$$p(y_{t+1} \mid h_t, u_t) = p(y_{t+1} \mid a_t, u_t). \tag{54}$$
3. The expected reward is conditionally independent of the observed sequence given the statistic of the sequence,
$$\mathrm{E}[R_t \mid h_t, u_t] = \mathrm{E}[R_t \mid a_t, u_t]. \tag{55}$$

Given these properties, infomax control can be expressed as a Markov decision process where the state is given by the statistic A_t. This allows for solving the Bellman equations using standard dynamic programming and reinforcement learning approaches.

Appendix D. Infomax TD learning

We used the following finite horizon version of value based TD(0) learning. For each state a_t of the A_t statistic, and for each time $t = 1, \ldots, \tau$, we initialize the value estimates $V_t(a_t)$ to zero, which is an optimistic value. Each learning trial starts at time $t = 1$ and ends at the terminal time τ. At time $t = 1$ we draw s, k_1, k_2, and k_3 from their prior distributions, and initialize a_1 to $\{Z = z_1, P_2 = Q_2 = P_3 = Q_3 = 0\}$, where z_1 is drawn from the uniform probability distribution over the range $1 : \tau_2^a + 1$. Then for $t = 2, \ldots, \tau$, we choose with probability $(1 - \epsilon)$ the action $\hat{u}_t$ that maximizes the expected value:

$$\hat{u}_t = \operatorname*{argmax}_{u_t} \sum_{y_{t+1}} p(y_{t+1} \mid a_t, u_t) V_{t+1}(a_{t+1}) \tag{56}$$

where

$$a_{t+1} = f_{t+1}(a_t, u_t, y_{t+1}). \tag{57}$$

With probability ϵ we choose the other action. After each trial, we perform backups to the value estimates $V_t(a_t)$ of each visited state a_t, in reverse order, according to the following equation:

$$V_t(a_t) = r_t + \sum_{y_{t+1}} p(y_{t+1} \mid a_t, u_t) V_{t+1}(a_{t+1}) \tag{58}$$

where

$$r_t = -\mathcal{H}(S \mid a_t). \tag{59}$$

For the terminal time τ we simply let

$$V_\tau(a_\tau) = r_\tau = -\mathcal{H}(S \mid a_\tau). \tag{60}$$

The update equations are repeated for multiple trials. As the number of trials increases, the estimate of the value function $V_t(a_t)$ converges to its true value. At evaluation time, setting ϵ to 0 gives the optimal policy.

References

Bertsekas, D. (2007). *Dynamic programming and optimal control*. Athena Scientific.

Bertsekas, D., & Shreve, S. (1996). *Stochastic optimal control*. Athena Scientific.

Bigelow, A. E. (1999). Infant's sensitivity to imperfect contingency in social interaction. In P. Rochat (Ed.), *Early social cognition: understanding others in the first months of life* (pp. 241–256). New York: LEA.

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–126.

Butko, N. J., & Movellan, J. R. (2008). I-POMDP: an infomax model of eye movement. In *Proceedings of the 2008 IEEE international conference on development and learning*.

Butko, Nicholas J., & Movellan, Javier R. (2009). Optimal scanning for faster object detection. In *Proc. IEEE conference on computer vision and pattern recognition*.

Butko, Nicholas J., & Movellan, Javier R. (2010). Infomax control of eye movements. *IEEE Transactions on Autonomous Mental Development*, *2*(2).

Edelman, S., & Vaina, L. M. (2001). David Marr. In *International encyclopedia of the social and behavioral sciences*.

Johnson, S., Slaughter, V., & Carey, B. (1998). Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Developmental Science*, *1*(2), 233–238.

Marr, David (1982). *Vision*. New York: Freeman.

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*(7010), 760–767.

Movellan, J. R., & Watson, J. S. (1987). Perception of directional attention. In *Infant behavior and development: abstracts of the 6th international conference on infant studies*.

Movellan, J. R., & Watson, J. S. (2002). The development of gaze following as a Bayesian systems identification problem. In *Proceedings of the international conference on development and learning*. IEEE.

Nelson, J. D., & Movellan, J. R. (2001). Active inference in concept induction. In T. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems*: *vol. 13* (pp. 45–51). Cambridge, Massachusetts: MIT Press.

Nelson, J. D., Tenenbaum, J. B., & Movellan, J. R. (2001). Active inference in concept learning. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 692–697).

Bahrick, L. R., & Watson, J. S. (1985). Detection of intermodal proprioceptive-visual contingency as a potential basis of self-perception in infancy. *Developmental Psychology*, *21*, 963–973.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593.

Sutton, R. S., & Barto, A. G. (1988). *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.

Watson, J. S. (1972). Smiling, cooing and the game. *Merrill-Palmer Quarterly*, *18*, 323–339.

Watson, J. S. (1979). The perception of contingency as a determinant of social responsiveness. In E. B. Thoman (Ed.), *Origins of the infant's social responsiveness* (pp. 33–64). New York: LEA.

Watson, J. S. (1985). Contingency perception in early social development. In T. M. Field, & N. A. Fox (Eds.), *Social perception in infants* (pp. 157–176). New Jersey: Ables.

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, *5*(6), 483–494.