ABSTRACT
A central issue to address in an attempt to build machines that learn new things by interacting with people is the ability to understand intentions: the ability to predict and make expectations about the intended goal state of an agent’s actions. By the age of four, children are undeniably able to reason about intentions. In this paper I present various findings from cognitive science and developmental psychology about mechanisms that precede a fully developed theory of mind and may in fact be instrumental to its development: eye recognition and tracking, low-level pattern recognition of the physical and temporal structure of action sequences, intermodal self-representation, efficiency expectations for goal-directed actions, attraction to others ‘like-me’, and an appetite for being imitated. I then address each of these from an implementation perspective as design guidelines for building a machine that is to understand intentions.

INTRODUCTION
In the effort to build machines with human-like intelligence, one approach is to gain insight and clues from human development. Hence, to build a machine that learns something new, this approach directs us to look at how humans learn something new.

There are essentially two ways in which people learn new things: either through their own experience or through interacting with another person. Learning on ones own generally involves a process of trial and error, in which one has a particular goal or end state in mind (Minsky, 2003). Without discounting the importance of this process in the grand scheme, it is quite inefficient unless you have luck or prior experience to get you going in the right direction. From an evolution point of view, when survival is the ultimate goal, other people (especially those that infants are presumably with most – parents) are likely to be the most efficient source of information. In developmental psychology this is described as parental scaffolding, in which the parent provides constraints or structure to facilitate the learning process of the child. Researchers in artificial intelligence envision this kind of social learning as an avenue towards machines that acquire new knowledge autonomously and increase in complexity without intervention from the designer (Breazal, 2002).

Fundamentally, learning something new from another person requires being able to decipher their actions and understand of their intentions. How exactly children come to understand intentions is an active field in cognitive science and developmental psychology. In the first half of this paper, I present findings and speculation from these fields concerning the mechanisms instrumental to the development of an ability to understand intentions. In the second half, I then address these mechanisms from an implementation perspective, proposing five design guidelines that will be necessary when building a machine that understands intentions.

MECHANISMS FOR UNDERSTANDING INTENTIONS
The ability to learn new goals requires seeing the world not merely as a series of events but in terms of goal-directed action, an endless sequence of means and ends. Learning something new then involves finding a way to an end state has not been seen before or perhaps learning a new path to a known end state. If this learning takes place while interacting with another individual, it requires the ability to make sense of their actions and understand of their intentions.

Intentions and intentionality are part of that set of commonsense terms of which everyone has an understanding, but it is hard to find a single definition that covers all of what we mean by these words. In this paper I use the term understanding intentions to refer to the ability to predict and make expectations about the intended goal state of an agent’s actions. This involves reasoning about the agent’s actions in terms of their goals, beliefs and desires. This reasoning about mental states is why understanding intentions and theory of mind are sometimes used interchangeably in the literature.

The false-belief task has since the 1980’s been a widely accepted litmus test for theory of mind (Lewis & Mitchell, 1994). There are two widely used versions of the false belief task: the unexpected transfer, and the deceptive box. In a typical unexpected transfer experiment, a child sees one experimenter come into the room, put chocolate in a cabinet, and then leave the room. Then a second experimenter comes into the room and moves the chocolate to a different cabinet. When the first experimenter returns to the room the child is asked where they will look for the chocolate. In a typical deceptive box experiment, the child is shown a smarties box and asked ‘what do you think is in here?’ and as expected replies ‘smarties!’ Then the child is shown that there are actually pencils inside, and is asked what another child will think is in the box. Children under four consistently fail these tasks, and around the age four they become able to reason
about another person’s beliefs being different than their own and consistently pass these tasks.

The great distinction between abilities of 3- and 4-year olds in the false-belief task leads many scientists to the conclusion that theory of mind develops in stages and something just changes about a child around their fourth birthday. Other scientists however are not satisfied with this conclusion and are seeking to understand what precedes the four-year-old mechanism of reasoning about beliefs and understanding intentions. This section details various findings about mechanisms that precede a fully developed theory of mind and may in fact be instrumental to its development. This group of findings is not exhaustive, but in particular, these are findings that I believe lead to concrete principles toward developing intentional understanding in machines.

Self-Representation and the ‘Like-Me’ Attraction

We can learn a lot about what children understand of intentions by looking at imitation. Some of the most seminal and extensive work on imitation comes from Meltzoff and Moore. Numerous studies show that when children imitate they are doing more than copying actions, they are copying the goal and can do so in various contexts and delayed circumstances, well before the age of four (Meltzoff, 1993) (Meltzoff, 1996).

Meltzoff et. al. hypothesize that the ‘commonsense psychology’ or theory of mind that children show undoubtedly by the age of five has roots in infants’ ability to recognize that certain other things are ‘like me’ and their predisposition to attending to those ‘like-me’ agents. One of their most important findings is that very young infants (a few hours old) can and do imitate facial expressions (tongue out, lips out, lips open). In this study, the results of which have been replicated numerous times, infants’ faces are videotaped as they observe an experimenter making various facial expressions. Adults are asked to identify what facial expressions they think the infants on the video are making (not being able to see the model they were presented). The results show that the facial expression of the model significantly increased the likelihood of the infant producing the same expression. Moreover, in a variation of the experiment they proved that it is not just a reflex behavior. Until this, imitation was assumed to develop later but these findings suggest some aspects must be innate. The fact that this is facial imitation (and the infants have surely never seen their own face) shows an innate cross-modal imitation ability, where body-movements-as-seen are successfully mapped onto body-movements-as-felt. Meltzoff and Moore term this innate self-representation the ‘active intermodal mapping’.

In an additional set of studies it was found that infants also have an attraction or appetite for being imitated. They prefer to attend to an adult who acts ‘just like me’ and are recognizing both temporal and structural equivalences of the imitation. The fact that infants enjoy and draw their attention to people imitating them is essentially why parental scaffolding works, some actions are selectively imitated more often and thus reinforced.

This work points out three important precursors to understanding intentions: an intermodal body map, recognition and attraction to others ‘like-me’, and an appetite for being imitated. Meltzoff and Gopnik tie these together nicely and characterize mutual imitation as children’s ‘tutorial in naïve psychology’ just as observing physical reactions is a ‘tutorial in naïve physics’.

Parsing Action Sequences Along Intentional Lines

One way to view the challenge of understanding intentions is as a perceptual problem. The brain is constantly bombarded with streams of sensory data that it must translate into something meaningful. Understanding intentions is a continual process, when observing an action sequence understanding intentions involves parsing that action stream into groups of actions that collectively represent an intended act. Baird and Baldwin designed a series of experiments to explore the relation between intentions and action parsing (Baird & Baldwin, 2001).

In the same way that we naturally parse speech along phoneme lines, Baird and Baldwin found that adults naturally parse action sequences along intentional lines. The experiment involved video sequences of a woman cleaning her kitchen. A group of subjects coded the video sequences, determining precisely where adults agreed about one intended act ending and another beginning. Incidentally, it is interesting that there was a high level of agreement among the coders. For the experiment, short tones were placed in the audio track according to intentional lines. Tones were placed either at the endpoint of an intentional act or at the midpoint. Then a second group of adults each watched four sequences of the video, two with endpoint tones and two with midpoint tones, and were asked to remember exactly when the beep occurred. The four sequences were played back without the tones and the subjects were asked to click a computer mouse exactly where they remembered the tone. If adults do indeed parse action sequences along intentional lines then one would expect it to be cognitively easier to do this task when the tones happened at an endpoint rather than a midpoint of an intentional action, and this is just what was found. The adults were significantly better at determining the placement of the tones at intention endpoints.

This is not entirely surprising that we automatically parse an action sequence into intentional acts, but leads to an interesting question: does this tendency come from an understanding of intentions, or is this something we do before having a fully developed understanding of intentions. Following their study of action parsing in adults, Baird and Baldwin next designed experiments to study whether or not there is evidence of this in infants.
Infants were shown the same sequences of kitchen cleaning video. Instead of placing tones in the sequences, the infants were shown some sequences in which the video paused at the endpoint of an intended act and other sequences in which the video paused in the midpoint. Using the ‘looking time’ metric that is widely accepted in child psychology to reveal an expectation violation, they found that infants looked much longer at the sequences in which the intended act was interrupted. This indicates that even without any high-level knowledge about cleaning kitchens the infants were still parsing the action sequences along intentional lines.

Given that infants show some ability to parse action sequences along intentional lines before showing any high level intentional reasoning, this suggests that there is perhaps a low-level mechanism at work. Baird and Baldwin call this a two-tier model of intention inference in which a high-level analysis works in parallel with a low-level analysis of the physical and temporal structure of action sequences. An example of what such a low-level domain-general structure might be: eye-gaze as a detector of accidental action. When there is a change in eye gaze that follows an action rather than precedes it this is a good indication that the agent did not expect the results. This work would indicate that there are a number of such low-level physical and temporal structures (probably related to eye gaze, head position, and body positions) that the infants are using when they are observing the kitchen cleaning video.

The Efficiency Principle of Goal-Directed Action
In the same spirit as the study of action parsing, Csibra investigates how infants recognize goal-directed behavior, (Premack, 1990) and others since suggest self-propelled motion indicates goal-directedness for children. Csibra goes further to explain that infants use an efficiency principle to attribute goal-directedness. This principle requires that behaviors directed to the same goal be adjusted in relation to the relevant aspects of the environment in which they occur. Additionally, perception of a behavioral adjustment that is a function of situational constraints may serve as a trigger condition for attributing goal-directedness to a behavior. This section details the evidence of children’s efficiency principle or ‘teleological stance’ (Csibra, 2003).

A series of experiments explain infants’ perception of goal-directed action in terms of using of the ‘teleological stance’. Infants were habituated to seeing a computer-animated ball that has to jump over a wall to reach another ball. They were then shown a second situation in which the wall was there but the ball was already on the other side of it, and in this case they were either shown that the ball jumps to reach the other one or that it just rolls directly to it. The infants looked longer when the ball did the now unnecessary jump before reaching the other ball, showing they had adjusted their expectations in light of the new physical context. Thus indicating that they do use the teleological stance to form expectations about what actions should reach a particular end state.

The teleological stance also gives a nice explanation of how infants selectively imitate a new action. In one imitation experiment, infants (14 months) were shown six different acts on different objects including one novel act, bending from the waist and banging a panel with one’s forehead (this was tested for novelty with a control group). The infants were only allowed to observe the experimenter playing with the objects, and then after one week they returned to the lab and were allowed to play with the same objects. 67% of the infants in the imitation condition exhibited the bending to touch the panel with their forehead, compared to 0% in the control group, significantly exhibiting imitation of this novel act (Meltzoff, 1996). In a variation of this experiment, Gergely et. al. modified the presentation of the novel act. As the experimenter bent down to touch the panel with her forehead, she wrapped a blanket around her shoulders, thus making her hands unavailable. It was found that only a minority of the infants imitated using the head to hit the panel in the hands-tied version, suggesting that they have a representation of the physical context in which they saw an action reach the goal and they assume that in the new physical context/constraints they are free to choose the most efficient way to hit the panel (Csibra, 2003). By interpreting the world with the teleological stance, infants are able to extract what is necessary about the action of the model in a particular context and adjust those actions given a new context.

Paying Attention to Attention
Another major component of understanding intentions is having an understanding of reference or what a person is attending to. Understanding that what a person’s actions are about can be an object or something in the environment. Children’s aptitude at this is readily apparent in the context of learning language, which of course happens before the critical 4-year old age of interest. The fact that children exhibit ‘fast mapping’ (learning a new word from very few examples of its use) and have relatively error-free word learning is evidence that instead of passively associating sounds and images the child somehow actively tries to figure out the meaning the adult had intended to express (Bloom, 2002).

Even preverbal infants show some signs of understanding referential action. Gaze following has been found in lab situations in children as young as 3 month olds, but an important point was that the gaze following does not happen unless there is a period of eye contact before hand. This points to a two-step process, where infants will follow a person’s gaze if first a communicative situation is established. It has been shown that such a situation can be
established in two ways: by making eye contact, or by contingently responding to the infant’s actions and vocalizations. This shows an early understanding of referential actions especially in terms of eye direction, which could lead to a higher-level understanding of referential intentions (Csibra, 2003).

There exists neurobiological support for the brain structures involved in recognizing eye contact and shared attention (Baron-Cohen & Ring, 1994). There is evidence that the Superior Temporal Sulcus (STS) has cells that react to eye gaze (particularly eyes pointing forward), and cells that react to individual faces and orientation of head. Thus it is likely that the STS encompasses the mechanisms that allow children to assess and attend to other people’s attention.

AN IMPLEMENTATION PERSPECTIVE

All of this evidence suggests that as we are setting out to build a machine that understanding intentions it is likely to need to implement the following:

- Attention tracking and attraction to eyes especially ‘eyes looking at me’.
- Low-level pattern recognition of the physical and temporal structure of action sequences for intention boundaries.
- Intermodal self-representation.
- An attraction to others ‘like-me’ and an appetite for being imitated.
- Efficiency expectations for goal-directed actions.

This section addresses each of these from a design and implementation perspective, proposing future work and recognizing work already in progress.

Following Attention

The first design guideline we can take from the literature is the attraction to eyes especially ‘eyes looking at me’. Recognizing eyes in a visual scene is technology that is already available in systems like Blue Eyes (Morimoto et al., 1998). The harder problem is then tracking these eyes and following the direction of gaze to find what the person is attending to. The visual interfaces research group of the MIT AI Lab is in the process of ongoing research into the problem of gaze following and attention tracking (http://www.ai.mit.edu/projects/vip/).

Low-level Action Sequence Parsing for Intentions

Baird and Baldwin list as future work (none yet published) the problem of specifying the low-level tier of their model of intention inference. The problem can be described as finding the domain-general physical and temporal qualities of action that covary with transitions between intentionnal actions. Given that they have a coded dataset of video sequences from their experiment, they are in a good position to make progress on this problem. With the labeled sequences, some feature detectors could be devised to extract physical features from the video. The physical features that are likely to be relevant are things like change in eye-gaze, head position, and speed of body movement; extracting these features from a visual scene may prove to be difficult. A cumbersome alternative would be to have another group of people label some physical features of the video. Given physical features and intentional boundaries, then standard supervised learning techniques can be applied to correlate the two. Additionally we learned from their experience that there was extremely high correlation about intentional boundaries among coders (i.e. this is an easy dataset to label); therefore, building ones own dataset and pursuing this independently is certainly reasonable.

The Intermodal Body Mapping

I see two interpretations of Meltzoff’s imitation findings. It is possible, as they suggest, that a complete intermodal body map is innate, in which all body-movements-seen can be mapped to body-movements-felt. However, it is also reasonable, that only the facial part of this body map is innate. Again going back to the perspective of evolution where survival is the key, being able to imitate facial expressions right away would be quite beneficial. From Ekman we have learned that facial expressions invoke a physiological response; just putting the face in a particular position can change a person’s affective bias (Ekman, 1983). A predisposition to imitate facial expressions gives the infant an innate ability to assess its environment through an adults facial expression, invoking fear in a dangerous situation or happiness in a safe situation, which makes a lot of sense. I am a bit skeptical of making the intellectual jump that Meltzoff’s facial imitation findings suggest a complete innate intermodal body map. This issue will be resolved as scientists continue to learn more about the development of imitation abilities.

Assuming that there is an innate body mapping as Meltzoff has suggested, let us turn to how this could be implemented. We have some neurobiological clues about intermodal mapping; mirror neurons seem a likely candidate for how this works. These neurons show activity both when a specific action is performed and when the action is seen performed by others. The current extent of our knowledge about the workings of mirror neuron systems is that they do exist, and it has been shown experimentally that they are reacting to whole motor gestures not just specific muscle movements. They appear to have the capacity to be a ‘supramodal representation’ connecting visual processing and the motor cortex (Williams et al., 2001).

This body-movements-seen to body-movements-felt mapping requires self-awareness: a representation of what one’s own body ‘looks like’, how it can move, and probably mechanisms to simulate self-movement for planning purposes. Then the see-to-do mapping is
achieved through matching this internal representation to a
visual perception of the actions of others.
Sensing and representing the global body position: I
propose that we could take advantage of the fact that a
body is made up of connected parts. If these parts have a
unique identifier, information about their physical
qualities, and can communicate to a central system about
their relative position to other parts; then this central
system could build a representation of the current body
state from the bottom up. The construction starts at the
ground, so the first step is to find all parts touching the
ground. Given their physical description, draw these parts.
Then find the connector linking this part to others, and ask
the connector which parts are connected at what angles.
Now draw these parts and recursively go through this
process until all body parts have been drawn. Now the
problem boils down to representing the physical qualities
of a single body part in such a way that the system could
draw a visual representation. This is still not a trivial task,
but what I like about this approach is that it could scale.
This self-representation builder should not need to be more
complicated for a robot with 100 body parts than for a
robot with 5 body parts. The biggest issue is probably
going to be the time it would take to build the
representation. Since this process would need to ‘redraw’
every time any part moved, it could potentially be too slow
to keep up with the activity.

Given a way to sense the global body position and have a
visual representation of the current body state, the robot
can do some experimentation to find out what its behaviors
‘look like’. This presents us with both body-movements-
seen and body-movements-felt as a visual representation:
an intermodal mapping.

Attention to ‘Like-Me’ Agents and Imitating Agents
A body-movements-seen to body-movements-felt mapping
enables the ability to recognize when a motion seen
represents an action that could be done. If this happens a
significant number of times then this agent can be
considered ‘like-me’. The attention system should be
biased to attend to these agents when recognized.

The intermodal body mapping also underlies the ability to
recognize being imitated. By remembering what you just
did, recognizing imitation becomes recognizing contingent
action that is similar to the actions just performed. The
literature also suggests that not only are children able to
detect when they are being imitated, but they prefer to
attend to those that are imitating them; therefore, the
attention system should be designed to give attention to an
agent detected as imitating.

Efficiency Expectation for Goal-Directed Action
There are two common explanations for how our theory of
mind works. Theory-Theory is the behaviorist point of
view, that we understand the intentions of others through
tacitly known causal laws, relating causal changes in
external stimuli to internal states. On the other hand, this
efficiency expectation is related to the Simulation-Theory
which proposes that we use or own mechanisms of
reasoning to simulate mental state of others (Gallese,
1998). In this approach the system would make
expectations about intended goal states using its own
planning system to simulating possible plans of action.
Given a planning system that has the ability to devise the
optimal path of actions from a current state to a goal state
in the current physical context, the efficiency principle
gives us two guidelines for forming expectations. In
making expectations about a future goal state, the system
should map the action seen to an action it can do and
determine the goal state for which that action is an
instrumental part of the plan. Additionally, when looking
for goal-directed behavior, the efficiency expectation
should be a trigger. The system can map the prior actions
seen to actions it can do and determine if this was an
efficient plan to reach the current state in the given
physical context.

Open Questions and Implicitly Hard Details
Memory is implicit in the idea of learning something new.
There needs to be a representation of the past, the ability to
recall a representation of a past event and compare it to
what is happening now, along with metrics of similarity for
these two events. I mostly have questions regarding this
issue, which I will list as thought experiments for the
reader. I believe the answers to these questions will be
discovered through implementation and experimentation.

- What is happening now can be very similar to a past
  perception or experience, but it is never exactly the
  same...where is the line drawn between similar and
  new?
- When a particular change in the world is seen what is
  significant and salient and thus important to save in a
  memory?
- Should multiple memories of the same event be saved
  as different points of view, or different levels of
  abstraction?
- How should a memory be triggered or recalled as
  being relevant to the current situation?
- Does something old in a new context count as
  something new?

Many of these questions come down to representation.
The choice of representation underlies learning new things
because the features of perception define the resolution
with which the system is going to be able to discriminate
and judge similarity. In implementing a system, this first
step of choosing a representation to work with inherently
limits the capabilities of the system. I believe that one of
the major roadblocks to a flexible learning machine is a
flexible representation, the ability to change one’s point of
view. This includes being able to represent past and
current events at many levels of abstraction and continually looking for the appropriate level of abstraction in which to be reasoning. Moreover, flexibility not only means being able to choose the right level of abstraction, but the ability to discover new ways to represent the world.

CONCLUSIONS
In any attempt to build a machine that can learn new things from people, the ability to decipher and understand intentions becomes a central issue. I have presented findings from cognitive science and developmental psychology that provide concrete guiding principles for building a machine that understands intentions. Understanding intentions is the ability to predict and make expectations about the intended goal state of an agent’s actions. A few precursors to a child’s ability to understand intentions include: eye recognition and tracking, low-level pattern recognition of the physical and temporal structure of action sequences, intermodal self-representation, efficiency expectations for goal-directed actions, attraction to others ‘like-me’, and an appetite for being imitated. Work is currently in progress on the issues of tracking eye gaze and identifying low-level action sequence features for intention boundaries. The remaining mechanisms rely heavily on our ability to implement an intermodal body map, and I have proposed a possible solution. From an implementation perspective, defining the appropriate representation and metrics of similarity for events in the world remains an open question. This will be instrumental to our realization of these intention precursor mechanisms and flexible machine learning in general.

REFERENCES