Robot's Play: Interactive Games With Sociable Machines

Andrew G. Brooks
Robotic Life Group
MIT Media Laboratory
77 Massachusetts Avenue
Cambridge, MA, USA
zoz@media.mit.edu

Jesse Gray
Robotic Life Group
MIT Media Laboratory
77 Massachusetts Avenue
Cambridge, MA, USA
jg@media.mit.edu

Guy Hoffman
Robotic Life Group
MIT Media Laboratory
77 Massachusetts Avenue
Cambridge, MA, USA
guy@media.mit.edu

ABSTRACT

Personal robots for human entertainment form a new class of computer-based entertainment that is beginning to become commercially and computationally practical. We expect a principal manifestation of their entertainment capabilities to be socially interactive game playing. We describe this form of gaming and summarize our current efforts in this direction on our lifelike, expressive, autonomous humanoid robot. Our focus is on teaching the robot via playful interaction using natural social gesture and language. We detail this in terms of two broad categories: teaching *as* play and teaching *with* play.

1. INTRODUCTION

Games have always been primary motivators of computer-generated entertainment. Indeed, game development and computer architecture have exhibited something akin to a bidirectional symbiosis. Increased computational power has led to aggressive incorporation of cutting-edge techniques within complicated games, such as artificially intelligent enemies and distributed networked gameplay; conversely, it is the popularity of games that has driven the development of affordable high-performance graphics hardware.

We are now witnessing sensor development being energized by new attention to the sensing of human behaviour for the purpose of game input. Some arcade games now feature ultrasonic tracking of the player's body to allow real-world ducking and weaving to be incorporated into gameplay; cheap computer vision systems and games to go along with them, such as Sony's EyeToy for PlayStation 2, are beginning to be marketed to home users; and novel sensors now in development, such as the ZMini 3D camera, have games as their intended market.

We expect the push towards games involving advances in the inclusion of real-world presence to now extend into the realm of mechanical actuation. We have already seen games bring force and vibrotactile feedback devices out of research laboratories and into the home; the logical ultimate extrapolation of real-world sensing and actuation for games is a game-playing robot.

Home entertainment robots have already emerged on the market in the form of robotic pets such as Sony's Aibo, and Sony has recently announced and demonstrated the walking humanoid QRIO targeted at the consumer market (though not yet actually for sale). While these will continue to be an expensive luxury item for some time, the price point can be expected to come down with further mass production and competition. However, at present these robots are minimally interactive or communicative, concentrating on replaying a set of simple behaviours and pre-scripted dance routines. The closest to game play is the QRIO's routine of tracking a coloured ball and kicking it.

Some games involving robots have entered the popular consciousness, but these have so far tended to involve robots playing against other robots according to formal rules, sometimes autonomously as in Robocup, and sometimes under human control as in the Robot Wars television show and the FIRST competition.

Some early intelligent robot work in the laboratory involved games, but typically of a highly formal and non-interactive nature, such as having a robot arm make physical chess moves during a game against the computer. For instance, the chess computer Deep Blue is able to compete against the world's best human chess player, but it is essentially immaterial whether the pieces are moved by a mechanical actuator or a human attendant.

Our aim is to make robots able to compellingly and autonomously entertain humans with social games. This paper introduces our arguments and theoretical framework for social game-play and our work towards implementing these on an expressive humanoid robot, Leonardo (Figure 1), with dual emphases on learning the game-play itself and entertaining the human teacher during the process.

2. THEORETICAL FRAMEWORK

In considering a robot for real-time face-to-face game interaction with a human player, we discuss motivations and core research problems. We describe the games and social responsiveness we have in mind, and discuss the application of game-play to the question of learning, taking inspiration

ACE 2004, Singapore

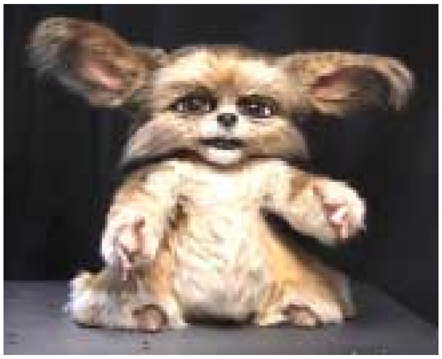


Figure 1: Leonardo, a lifelike robot designed for social interaction.

from human imitation and simulation theory.

2.1 Game Playing Robots

There are now many examples of the use of agent-based technology to simulate characters for entertainment purposes. Artificial creatures with individual behaviour models are now routinely used to create realistic visual flock behaviour for animation and movies (e.g. [28]). Agents with natural language processing abilities have been used to populate online chat rooms and multi-user worlds with characters that interact in a human conversational fashion (e.g. [18]). And full-body, face-to-face interactions with synthetic characters have begun to be investigated, though typically with animated versions rather than physical robots. For example, Blumberg et al. have developed a number of animal characters with interaction-based motivations for immersive play (e.g. [17]). The robot Kismet was designed for emotionally engaging interaction in a face-to-face fashion [3]. However, these experiments have not focused on gaming as the primary interaction model.

In the majority of cases of computer-based agents designed specifically for playing games with humans, the level of entertainment is strongly dependent on the human: the human will have fun if and only if he or she is predisposed to the game. To some extent this is clearly inevitable, but we believe that embodied, socially aware characters can have the ability to improve the experience for the human player. We therefore do not consider games in which it is not important whether the non-human player has a physical or virtual presence, such as formal games like backgammon or checkers. The motivation in the latter case is simple: to win the game according to strict pre-existing rules, and it is not necessary for the computer to adapt to the playful attitude of the human.

Similarly, we consider a different type of game than that in which the most important aspect of the players' presence is their movements in space and time, and the nature of the interaction is restricted and secondary. The computer's motivations in this case may be more varied, but the ultimate goal tends to be to play at a certain level of skill, leaving it up to the human to mediate the entertainment by selecting the level of preference; some like a challenge, some to always win. Most related work in synthetic game-playing characters to date has been of this type, such as the development of artificially intelligent "bots" in multiplayer online environments such as Quake, or of real robots that play soccer against humans [24, 8]. In these cases their operation is completely autonomous and the nature of the interaction is highly restricted – to kill the enemy players, or to score goals.

For a game in which a direct, primary goal is the satisfaction of the players, a robot needs to know more about how humans have fun, and to be able to determine whether they are having it or not. Social and affective competence of the robot is required to infer whether people are entertained and to respond appropriately to keep them engaged in the game.

Finally, in our research we concentrate on real robots with a physical presence. Physical presence seems to relate to social presence, and there is a strong argument that due to the limitations of display technology a physical robot is more engaging of the human's senses. In restricted scenarios in which humans and artificial characters interact one-on-one, humans were observed to prefer a real robot to an animated version [16]. We take this as an indication that humans would derive particular enjoyment from playing socially interactive games with robots rather than just graphical characters on a video screen. This adds complexity to our research, such as the need to develop control systems for lifelike, expressive motion (see [5] for related work in this area). But we believe that the difficulties are surmountable and that doing so will ultimately prove rewarding in terms of the compelling nature of the resulting game interaction.

2.2 Social and Affective Response

As with other interactions, it makes design sense to craft human-robot game interactions in terms of people's pre-existing social model. Humans are experts at social interaction and thus in most cases are already both familiar and comfortable with the interface [27]. It has been shown that people will apply a social model when interacting with a sociable robot [4], so we expect this to extend to sociable game playing robots. The types of social games that we have in mind for human-robot interaction purposes are games that are cooperative and highly contingent (rather than each player proceeding completely autonomously, as in games such as Quake or soccer). Turn-taking must be recognized when necessary, and each player's behaviour must be dependent on that of its counterpart.

A simple example is the game of poker. Poker has formal rules and is easy to program a computer to play, but to play poker properly, as between humans, demands a high level of social inference that requires a detailed bodily presence. However, poker does have a strict "winning condition" – to gain money. We are also concerned with games that have no strict way to win; instead, the object of the exercise is primarily to "have fun". The social aspect of the game is an important part of its entertainment – the fact that you play it with others, rather than the game itself. Rudimentary examples include many children's games such as hopscotch and patty-cake, as well as animal games such as "fetch" that might be implemented on a robotic pet.

If the overriding goal is for the human to have fun, we must develop representations and mechanisms to best insure that is the case. Approaching this from the standpoint of the design of the game itself — designing an attractive game that we predict will be entertaining — is only part of the equation. Since we are also designing a socially aware synthetic player, it is possible for us to incorporate heuristics for this purpose from a priori knowledge of what humans tend to enjoy from a playmate. For example, we might equip a game-playing robot with a sequence habituation mechanism to try to forestall boredom by limiting the amount of repetition in the game. Many games, however, such as video games, involve highly repetitious actions yet are still found enjoyable by human players. We must therefore also implement tight monitoring of and feedback from the human's affective response, to enable the robot to react to the individual player's immediate mental state. A player who currently does not feel like fighting can never have fun playing against current QuakeBots. But what if it were to do something unexpected, like juggle its weapons, when the player looked bored?

A core research issue in the design of robots for socially interactive game-play is thus appropriate detection and response to external manifestations of the human's affective state. This incorporates issues of both sensing and internal representation. There are obvious examples of instantaneous facial and audible responses — positive ones such as smiling or laughing, and negative ones such as frowning or crying. However, exaggerated responses cannot necessarily be counted on in a game context, even (or in some cases, especially) when both players are human. We suggest that a longer-term measure of approval would be a more useful feedback signal. For a face-to-face interaction, one such metric is the level of attentiveness and engagement in the interaction. This has been inferred by such means as eye contact and blink rate, posture, voice inflection, and blood flow to the skin (e.g. [12, 23]). For the moment we are concentrating on visual sensing through facial contact, though posture will be an important future target.

2.3 Imitation and Simulation Theory

In order to do the things we have mentioned, such as infer the affective state of the human player from observations of extended behaviour rather just overt emotional gestures, the game-playing robot requires a model with which to represent the human's mental state. In the psychological literature, such a model is referred to as a theory of mind [26]. In addition to the attentiveness and emotional state of the human, it may be useful to also keep track of what the human believes about the state of the game that might not be verifiable by immediate observation, and what the robot believes about the game that the human may not know. Taking the example of poker, this is clearly crucial to effective bluffing. A theory of mind allows a player to formulate more appropriate responses to developments in the game and actions performed by the other player.

One of the dominant hypotheses concerning the nature of the cognitive processes that underlie theory of mind is Simulation Theory [10, 13, 14]. According to this theory, humans are able to make assumptions and predictions about the behaviours and mental states of other agents by simulating the observed stimuli and actions of those agents through one's own behavioural and stimulus processing mechanisms. In other words, we make hypotheses about what another person must be thinking, feeling or desiring by using our own cognitive systems to think "as if" we were that other person. In addition to being well supported, Simulation Theory is attractive as a basis for the design of the mental architecture of a game robot because it does not require different models for the mental state of the human player. Instead, the robot could represent other people from instantiations of its own cognitive system, assuming that the robot is indeed enough "like" the human through body and psychology.

We are therefore interested in imitation learning and imitative games because there is strong evidence that imitation is one of the important bootstrapping factors for teaching infants this ability to simulate others [19, 20]. Games that involve human imitation are a natural way to provide entertainment that is tailored to the individual, as well as simplifying and making more enjoyable the teaching process. It may also bring us fundamental knowledge about how to design intelligent systems for the purpose of entertaining humans.

For example, much of the primate developmental psychology literature concerns learning from demonstrations in which a monolithic, tangible reward is present as a goal state. The representation becomes less clear when the more nebulous state of entertainment becomes the overall process goal. The schema that we find most useful breaks learning from social demonstration into three "information sources": goals, actions and results; and then uses combinations of these to produce a taxonomy of imitation processes such as mimicry, imitation and goal emulation [7].

When constructing a robot to learn from human demonstration, it may be useful to think of these processes as distinct learning modules. However, some modifications need to be made for the case of game-play. In this instance the robot may perceive directly conflicting goals even assuming that it has correctly isolated the demonstration. We separate these into two categories. First, there are "game goals", such as improving one's own performance towards the winning condition (if any), performing an action correctly, learning a sequence or taking one's turn when appropriate. Second, there are "entertainment goals", focusing on heeping the human engaged and entertained in various ways, such as parodying the other player, teasing, or pretending. These are likely to manifest themselves as actions that do not progress towards the game goals, such as reversing the other player's actions or deliberately underperforming a demonstrated step, for entertainment ends such as creating a humorous interaction. We therefore argue for a greater focus on cases in which the robot understands but chooses not to adopt the demonstration goal, cases which would from external observation tend to be described as simple mimicry or the absence of social learning, but which take new relevance in the context of a game.

We therefore present two directions of work that we are undertaking towards teaching a robot through imitative games that provide entertainment both during the teaching phase and during the playing of the game. Ultimately, we hope this work will allow us to separate game goals from entertainment goals beneath an imitative, interactive framework and allow us to develop robots that play with and learn from us in more enjoyable ways. First, we explore teaching *as* play, in which the human teaches the robot a game so that it can subsequently be played; and second, teaching *with* play, in which we concentrate on improving the robot's basic skills in a fashion that is enjoyable for the teacher.

3. EXPERIMENTAL PLATFORM

The physical platform for our research is Leonardo ("Leo", Figure 1), a humanoid robot with 65 degrees of freedom that has been specifically designed for social interaction using facial expressions and lifelike body poses. Currently, Leo does not speak and therefore relies on gestures and facial expressions for social communication. The robot's underlying software architecture consists of the following subsystems: speech recognition and parsing, vision and attention, cognition and behavior, and motor control. All of these can also be used to control "Virtual Leo", a graphical simulator for the real robot.

3.1 Perceptual Systems

The robot has both visual and speech inputs. The visual system consists of cameras within the robot's mobile eyes, a stereo camera behind the robot's head and an overhead stereo camera pointing down into the robot's interaction space. The vision system concentrates on the detection of humans and specific items within the visual scene, such as objects with which the robot can interact. These perceptions are sent to the cognitive system along with object attributes (e.g., color, location). The vision system also recognizes pointing gestures and uses spatial reasoning to associate these gestures with their object referent. The speech understanding system is a Lisp parser based on the NRL Nautilus project [25] with a ViaVoice front end. The system has a limited grammar to facilitate accuracy of the voice recognition. Upon receiving phrases from ViaVoice, the speech understanding system parses these into commands that are sent to the cognitive system.

3.2 Cognitive System

The cognitive system extends the `C5M` architecture, developed from the `C4` system described in [2]. It receives a continuous stream of symbols from the vision and speech understanding systems and matches these against a tree-structured detector of perceptual features, or "percepts". Triggered percepts are recorded as beliefs about the world. The perceptual attributes of a given object are merged together and kept in one structure. For example, as shown in Figure 3, Leonardo's world is currently populated with buttons he can press. Information about a button's features such as location, color and ON/OFF state are merged to form a coherent belief about that button. These belief structures can also be manipulated internally, allowing the cognitive system to add information to its beliefs about the objects in the world (e.g., associating a label with a particular object so that the human can refer to it by name).

4. TEACHING ROBOTS AS PLAY

Teaching for entertainment purposes has long been a fundamental component of cooperative play. Both the process and the results are enjoyable — the fun is in teaching the game, and then in playing it together afterwards. Childhood games, such as patty-cake, often incorporate a teaching phase in most instances of the game, even though the game structure itself is easy to grasp. Similarly, people derive a great deal of enjoyment from teaching games and tricks to their pets, in addition to that gained when reproducing them afterwards on command. This has now extended to the cybernetic domain, with entertainment products incorporating teaching elements, such as Electronic Arts' game Black & White and PF Magic's synthetic pet software Petz. The nascent realm of home entertainment robots has been slow to follow, however — while code hackers are able to add new programmed behaviors to their Aibo, there are few facilities for teaching them in the conventional sense. There has been "clicker" training done on Aibo at the Sony CSL lab in Paris to teach Aibo new tricks, but not games as is our work.

We have been taking the first steps to enable real-time, face-to-face teaching of playful interactions to our humanoid robot Leo. The human and robot communicate via visually recognized body gestures that are natural (in the case of the human) and lifelike (in the case of the robot), as well as speech from the human (Leo can not yet reply in kind). The game interface involves a set of three buttons in front of the robot, that can be pressed ON or OFF by either the human or the robot. A change in the button's state is observable by an LED in the button which is switched on or off to reflect the relevant condition. An example of a simple game that can be taught and played with these buttons include human and robot competing to reach a desired game state (e.g. all buttons ON or OFF); a more complex version could have the buttons make different sounds when pressed, and the goal is to teach the robot to cooperate to play a tune.

4.1 Gestural Interaction

We wish for humans to be able to utilize their natural skills at social expression and understanding when interacting with Leo. We therefore use unencumbering (i.e. without marking or instrumenting the human) computer vision techniques to recognize bodily gestures on the part of the human, and natural language processing for recognizing speech. As Leo currently cannot speak, his means of communicating back are through the use of expressive body posture and facial expressions.

A multi-person visual tracking system has been implemented which can segment people from the environment and track the position of their bodies and heads. Detection of humans accomplished by background subtraction in both the video intensity and stereo-derived depth domains. Background subtraction in the intensity domain is a common computer vision technique for distinguishing the salient features of an environment based on the fact that they move, but may not be continuously in motion. Being able to combine these results with those from the stereo computation results in a significantly more robust detection image than referring to the intensity domain alone [9, 11, 15]. The person tracker then builds a depth histogram of foreground objects that allows the surface area and shape of each object in the scene to be calculated and compared to a template of a generic human. Regions that match are analyzed with a Viola-Jones face detector to determine if the person is facing the robot [29].

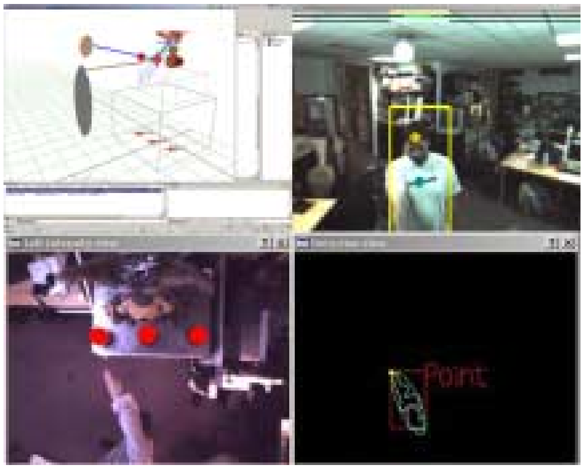


Figure 2: World model representation and human tracking and gesturing output.

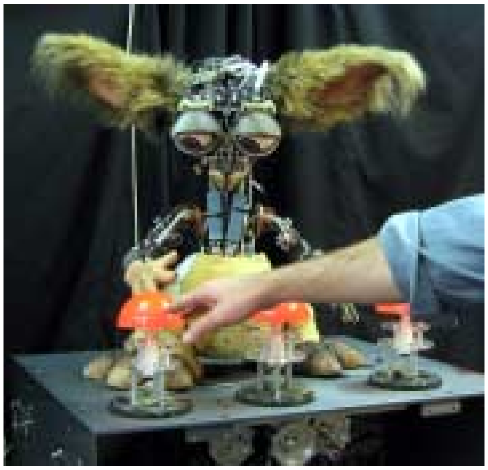
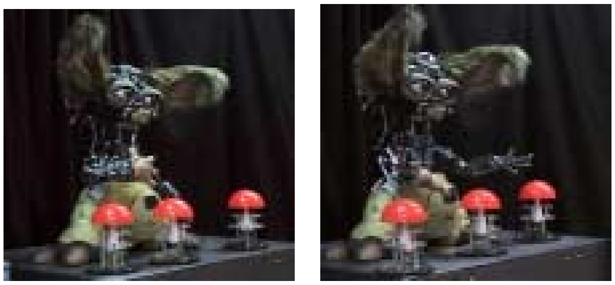


Figure 3: The robot acknowledges the human's gesture towards a button by looking at that button.



(a) Leo indicates willingness by gesturing towards himself.

(b) Leo asks for help by gesturing towards the human player.

Figure 4: Communicative gestures from the robot.

If not, a blob segmenting algorithm is run on the head region, assumed to be the top of the human object. Brightly colored objects such as the game buttons are detected and tracked via saturation matching in the HSV color space.

Our method of choice for establishing object reference from the human to the robot is the pointing gesture. Detection of the human arm is accomplished by background subtraction the video intensity and stereo depth domains provided by the overhead camera system. The largest separate candidate regions are then fit with a bounding ellipse from image moments within each region, and evaluated for likelihood of correspondence to an arm based on orientation and aspect ratio. The best candidate passing these tests is designated to be the pointing arm. Once the arm has been extracted, the task of recognizing whether the hand is configured in a pointing gesture or not is accomplished by estimating the kurtosis, or "pointiness", of the hand. Starting from the most distal point of the hand in the direction of the gross arm orientation, the deviation of the hand outline from the centreline is summed and thresholded to give a "point" score, voted on by several consecutive frames for robustness. When a pointing gesture is detected, the destination object is computed by a spatial reasoning system that extrapolates the gesture to the most likely destination in Leo's visual world model, which is stocked with the results of all of the visual detectors. Example detection outputs from all of these systems can be seen in Figure 2.

4.2 Collaborative Dialog

We model teaching and learning as a fundamentally collaborative process, in which it is important to open a channel of communication back from the learner to the teacher. When Leo recognizes that a gesture or an instruction has been made, he communicates back his understanding or lack thereof. At the start of the learning process, he indicates that he does not know a requested action by shrugging his shoulders and making a confused facial expression. As he is walked through each task in the game by the human, Leo offers visual feedback — for example, when the human partner changes the state of the world, Leo acknowledges this by glancing briefly towards the area of change before redirecting his gaze to the human, reassuring the instructor that the robot is aware that the action has been done (Figure 3). Similarly, Leo produces subtle nods while looking at the human to indicate when he thinks he has completed a goal. Communication such as this is crucial in allowing the establishment of mutual beliefs on the progress of the interaction.

In addition, Leo uses gesture to communicate his own capabilities. If he is able to perform an action, he points to himself and adopts an alert posture and facial expression (Figure 4(a)). Conversely, if he cannot perform it and wishes to ask for help, his espression indicates helplessness as he gestures towards the human in a request for a demonstration of the intended action (Figure 4(b)). In concert with this gesture, Leo shifts his gaze between the object and the human to direct the human's attention to what it is that he needs help with. Visual feedback from object and face detectors running on Leo's eye cameras compute the error between his gaze point and the center of the desired target, allowing him to look directly at it in a natural and believable flowing motion. Humans are easily able to perceive errors of even a few degrees in eye contact, so this ensures Leo acts as should be expected from a socially aware play partner.

4.3 Goal Representation

We suggest that an appropriate way to represent activities in a game is in terms of goals, rather than specific actions or motions. Goals provide a general common ground for participation in a shared activity, whereas actions may be sensitive to context that has not been represented. In the current case both the goals and the actions necessary to satisfy those goals are taught to the robot by demonstration, so in terms of Call & Carpenter's taxonomy the process being undergone is strict imitation. However, this goal-centric approach supports a more realistic groundwork for intentional understanding — i.e., to perform the task in a way that accomplishes the overall intent, rather than just mechanically going through the motions of performing the constituent actions.

To support this idea, we have extended the notion of the `C5M` *action-tuple* data structure. An action-tuple is a set of preconditions, executables, and until-conditions [2]. Both tasks and actions are represented as variants of this action-tuple structure, with the added notion of goals. As the robot learns a new task, it must learn the goals associated with each action, each sub-task, and the overall task. Goals are currently categorized as two distinct types: (a) `state-change` goals that represent a desired change in the observable state of the world, and (b) `just-do-it` goals that must be executed regardless of the state of the world. The type of goal affects the evaluation of both preconditions and until-conditions of action-tuples. As part of a precondition, a `state-change` goal must be evaluated before doing the activity to determine if the activity must be performed. As an until-condition, the robot persists in trying to perform the action, making multiple attempts if necessary, until it is successful in bringing about the desired state change. In comparison, a `just-do-it` goal will always lead to an action, and will always be performed only once.

Tasks are represented as a hierarchical structure of actions and sub-tasks, defined recursively in the same facshion. When learning a task, a goal is associated with the overall task in addition to each of the constituent actions. Overall task and sub-task goals are distinct from the mere conjunction of the goals of their actions and sub-tasks, and are learned separately. When executing a task, goals as preconditions and until-conditions of actions or sub-tasks manage the flow of decision-making throughout the task execution process. Overall task goals are evaluated separately from their constituent action goals to determine whether they need to be executed or whether they have been completed.

The goals currently taught to the robot come under the category of game goals. It is clear that this structure could also be used to represent entertainment goals once we are adept at discriminating them from the teaching process. This structure also does not forbid the incorporation of an innate playfulness into our robot character. When the human's emotional state is accurately reflected in the perceptual world model, `state-change` goals will be able to refer to changes in this emotional state also. One can imagine the robot performing unexpected actions to recapture a bored partner's attention, or trying multiple times to trigger a smile or laugh.

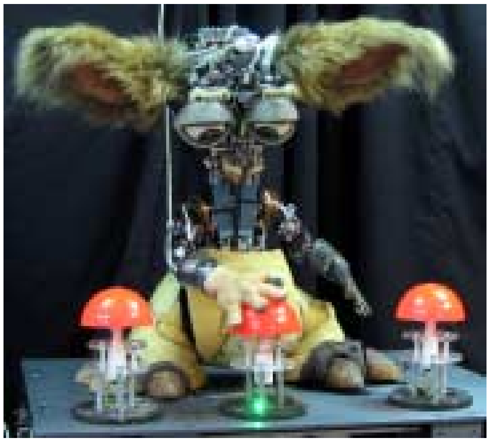


Figure 5: Leonardo performs the steps as he learns them, allowing the human teacher to catch errors in real time.

4.4 Game Learning and Play

Leo begins the interaction equipped with his repertoire of social responses, but not knowing how to play. If he is asked to do a task that he does not know how to perform, as he certainly will be, a learning module is instantiated. As the human teacher leads him through the task, the use of sequencing words naturally indicates the possible constraints between task steps. Since Leo shows his understanding of a newly learned sub-task or action by actually performing it (Figure 5), failure to comprehend an action or its goal is easily and naturally detected by the human player.

While in this task learning mode, the learning module continually pays attention to what actions are being performed, encoding the inferred goals with these actions. When encoding the goal state of a performed action or task, Leo compares the world state before and after its execution. In the case that this action or task caused a change of state, this change is taken to be the goal, of type `state-change`. Otherwise, the goal is assumed to be of the `just-do-it` type. This produces the desired hierarchical task representation, where a goal is encoded for each individual part of the task as well as for the overall task. When the human indicates that the task is done, it is added to the collection of known tasks. The process is performed recursively for tasks which involve sub-tasks.

This method of goal classification makes some arbitrary, though reasonable, assumptions as to the priority of world state over action. However, once the perceptual support is in place to allow the incorporation of entertainment goals, there will be questions of priority among the different facets of world state, as well as instances in which it might be possible to predict a need for action over state change. We are thus currently working on a more flexible method to learn the general goal structure.

Once Leo knows how to do things, he can be requested to do them by the human player. When this occurs, a collaboration module is started for that activity. This allows Leo to perform the task while allowing for the participation of the human player. Leo derives his plan of action based on a dynamic meshing of sub-plans according to his own actions and abilities, the actions of the human player, his understanding of the goal, and his assessment of the current state. At each stage of the game, either the human or Leo should take action. Before attempting the task element, Leo negotiates who should complete it. While usually conforming to this turn-taking approach, our system also supports simultaneous action, in which the human performs an action asynchronously. If this is the case, Leo will re-evaluate the goal state when he finishes his current action, and decide whether or not further action is required. This ensures that the intended activity is accomplished, rather than simply executing a sequence of planned actions without considering their consequence.

In our example scenario, we teach Leo tasks that involve turning his buttons ON and OFF by pressing them. The task set can represent both simple and complex hierarchies, and can have tasks with both state-change and just-do-it goals. For example, turning a button ON can be a single action, or a sub-task of turning all the buttons ON. Games can be developed from this task set by setting appropriate goals for the human player once Leonardo knows how to play, and can proceed with or without explicit turn-taking. For example, a "puzzle" game with turn taking can be developed by teaching Leo that the goal is to reach a particular button configuration regardless of the initial world state. During the human's turn, he or she rearranges the current button state as desired; during Leo's turn, he works out how to return the buttons to his goal state, and does so. Alternatively, a game can involve simultaneous action on the part of both players. Leo can be taught that his goal is to keep all the buttons ON; as he tries to do so, the human player simultaneously exercises his or own goal of turning them OFF. Leo will notice the continuous changes in world state caused by the human player and react accordingly.

We are therefore currently able to play with Leonardo by teaching him rudimentary games. The entertainment from these games comes from the teaching process itself, the social interplay with the robot, and from executing the games once taught. However, the robot does not yet attempt to analyze the enjoyment of the game, deviate from the game rules, or otherwise respond to the affective state of the human player. In section 6 we discuss such important future work that must be done towards these fundamental social play issues.

5. TEACHING ROBOTS WITH PLAY

Play has long been recognized as an important aspect of animal learning, assisting in the development of abilities ranging from basic motor coordination to high-level behaviors. For example, cats learn to stalk prey through play, a tendency which is believed to be innate [1]. Developing humans demonstrate a strong affinity for play, and educational games are used to keep children entertained as they learn. It seems intuitively plausible that intelligent robots might be designed to learn from playful interactions also. Furthermore, in the case of robots designed for entertainment, it is advantageous to consider the enjoyment of the human instructor as well, as repetitive instruction through reinforcement can require a number of iterations that would easily become onerous if not designed with entertainment in mind. We believe that teaching a robot through game-play is a promising approach to both issues, and have thus applied this principle to our expressive humanoid robot. We refer to this class of learning game as teaching *with* play — the goal is to improve the robot's basic skill set, independent of the game itself, while keeping the teacher amused.

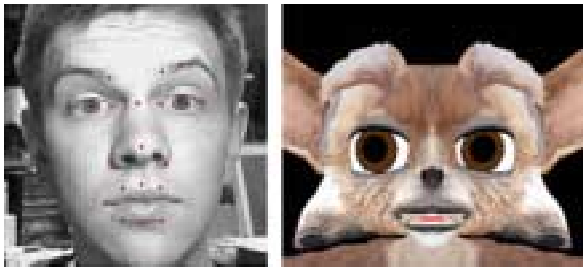


Figure 6: Facial tracking data is captured in real time in the form of the positions of 22 node points, and mapped to the pose space of an animated model of the robot.

5.1 Learning One's Own Body from Others

One skill we consider essential for future work is the ability of the robot to observe humans and understand their pose and movement relative to its own body. This enables the robot to physically imitate the human, which is as much an important ability as imitating the human in terms of goals and results. Recognizing other humans as "like me" and being able to physically map their bodies onto one's own is also believed by researchers to be part of developing the ability to simulate others in children [20]. In order to accomplish this task, the robot must be able to convert data it perceives about the human into its own joint space. In our case, Leo perceives human motion in two domains: facial motions and whole-body movement.

We receive data about the human's face as the 2-dimensional coordinates of various facial features, using a system based on the facial node tracker Axiom ffT (Figure 6). Body movement data comes from a motion capture suit worn by the human, with a joint structure similar but *not* identical to Leonardo's skeleton 7. The data that drive the robot's own motion is the rotation of various actuators which deform his skin or move his limbs.

Traditional machine learning solutions to this problem exist, but are time consuming and tedious to apply. Reinforcement learning is one option; the robot could attempt to imitate the human, and the human could reward correct behavior. A domain like this, however, has an extensive search space, and constant human supervision would be required to detect correct results. Also, there is the potential to damage the robot in the early stages of the process, as it tries to imitate the human with a flawed body mapping. Many supervised learning algorithms exist which use labeled training data to generate a mapping from one space to another (e.g. Neural Networks, Radial Basis Functions). These algorithms tend to require a large number of examples to train an accurate mapping, and the examples must be manually labeled. To apply such algorithms to our problem, we would require many examples of robot and human poses which have been matched together such that each pair contains robot pose

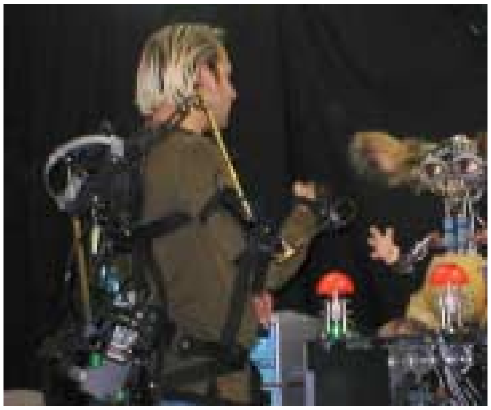


Figure 7: A motion capture suit captures human body poses which the robot must learn to relate according to its own body structure.

data that matches the associated human pose data. However, acquiring and labeling that training data through typical means would be quite tedious and time-consuming. We endeavor to transform that tedious interaction into an intuitive game played with the robot.

5.2 The Imitation Game

We have structured the game and the mapping process to streamline the teaching interaction for the human, making it easy and natural. The core of our mapping is based on Neural Networks for the face, and Radial Basis Functions for the body; however, the game is designed so that the robot can self-label the examples acquired during the game, removing the need for a manual labeling stage. The structure of the interaction also allows the robot to eliminate noisy examples immediately, requiring fewer total examples to learn the mapping. Finally, a special filter is applied to the output of the final map, allowing even a relatively under-trained mapping to produce good results.

In order to acquire the necessary data to teach the robot the mapping its face and body, we developed a simple but engaging "do as I do" game similar to that played between infants and their caregivers to refine their body mappings [22]. Acquiring bodily data in this way may then be taking advantage of an innate desire humans have to participate in imitative games. An additional benefit of this technique for a lifelike entertainment robot that might ultimately be destined for consumer or child use is that the entire process takes place within the context of social interaction between the human and robot, instead of forcing the robot offline for a manual calibration procedure which breaks the illusion of life.

Our imitation interaction is very simple: the robot motor system babbles through its repertoire of basis poses, chosen ahead of time to provide good coverage over all likely joint configurations, and the human attempts to imitate the robot. Because the human is actively imitating the robot, there is no need to manually label any data; each example saved should have the robot and the human in the same pose. In theory, since the human is intending to imitate the robot, any instant could provide a good teaching example. However, we found in initial tests that it was difficult and unnatural to imitate the whole robot continuously; teachers took breaks and often imitated only one limb or organ of the robot at a time.

To make the interaction easy for the human and at the same time ensure that we get the best data possible, we need to determine when the human is playing the game and what part of the robot the human is focusing on. It is difficult to determine with full confidence whether the human is imitating the robot before the robot has learned the mapping which relates the human's poses to its own, but certain heuristics can help with this problem.

It has been reported that human parents instinctively imitate their infants [19], and it is hypothesized that this information is used by the infant to solve the same problem we are addressing here: creating a mapping from perception to production. It is our belief that in learning tasks involving a human teacher there is supportive structure, or *scaffolding*, provided by the adult which helps the child learn by eliminating distractions or reducing the degrees of freedom involved in the task [30]. It can be advantageous to make use of this structure when designing a robot that will learn from a human. In order to begin to utilize this extra information, we have looked at the heuristics used to jump-start the development of imitation in human infants. For example, infants have been shown to respond preferentially to temporally contingent movements by caregivers, and it is believed that the infants use this information to detect when a human is imitating them [21]. Further, it is likely that the infants have some innate rough mapping that allows them to match their organs against the organs of the observed caregiver, even before they can necessarily imitate movements of that organ [22].

In order to detect this temporal contingency, the robot builds a model for the movement of each observable feature and determines when it moves relatively quickly and when it is still. This information can be compared with its own recent movement patterns, and it can guess that whenever the human's quick movements follow immediately after its own (the human's movements seem contingent on its own), the human may be currently imitating the robot.

Before being taught, Leo is not able to determine exactly how observed movements map to his own joint space; however, like the infants, the features can be divided into rough groups. For example, we can assume that only data obtained from the lip tracking features can control the mouth position of the robot, but we do not initially know how the 2-D lip positions map onto rotations of the robot's mouth area motors. This allows us to teach the mapping for each body part independently (lowering the number of examples needed), and also to apply the time-based metric to each body part individually to usefully use data from situations where the human is imitating some subset of the body parts of the robot.

Once the robot has acquired enough examples to create a mapping, it can map observed human poses into its own motor space. The robot does not use the output of this mapping directly; instead, it searches its basis set of poses (the same ones used for the teaching phase) to find the pose, or blend of poses, which most closely matches the observed human position. This may seem redundant, but it can significantly improve the robot's motor output if the data from the mapping is noisy. The basis set of poses is carefully chosen to span the set of possible positions of the robot, and resulting blends of these poses are lifelike and not dangerous for the robot. We found this to be especially noticeable in a complicated area like the mouth, where a mapping error even in one joint can cause an unnatural expression on the robot; after converting the mapping to a blend of basis poses, however, the robot cannot fail to produce something that is a natural blend of its basis poses (such as smile, frown, mouth open, etc.). This final step also has the advantage of producing a higher-level description of the human's motion; instead of a set of joint angles over time, the motion is now categorized in terms of Leo's own possible movements.

This technique has successfully resulted in a game which is easy and natural for the human, but still produces an accurate mapping from observed poses onto the robot's own motor space. We have used the system to allow Leo to imitate novel facial expressions [6], and found that the mapping created by one person can often be used by other people without having to teach the robot again. The final step of the mapping, finding Leo's own movement or blend of movements that most closely approximates the human's, has a number of benefits for our future work. Representing the human's facial expression as a weighted blend of Leo's known expressions allows him to try to categorize the expression and even begin to use social referencing. Also, knowing which of his own movements corresponds to the human's will be the basis of a simulation theoretic goal inference system for Leo. For example, if Leo notices a human performing a pushing motion near a button, he might detect the similarity to his own button pushing motion, then infer that the human's goal may be the same as his when he performs that motion near a button (i.e., activating the button).

6. FUTURE WORK

That robots are emerging from factories and science fiction and entering into the realm of computer entertainment devices is already a reality. This work represents the beginning of a long term investigation of how we can make robots more fun to interact, play and collaborate with by having them engage us better and learn from us.

Much remains to be done on the topics we have covered here, teaching a robot game-play and teaching it with gameplay. One direction of future effort will be on a more flexible structure of the goals that the robot can learn from human demonstration. We would like the robot to be able to make a decision on goal type precedence based on the context of the goal and the task. And of course we would like to be able to incorporate and detect entertainment goals. Two avenues of future effort will be creating innate playful behaviours for the robot, that allow it to have a greater repertoire of actions it can take to amuse the human player without detracting from its game goal progress, and in recognition of human behaviour that will allow it to identify when the other player is teasing it or pretending.

We also plan to add more perceptual skills to allow the robot to analyze the engagement of the human partner. In particular, we are in the process of incorporating a head pose tracker to provide the ability to extimate the human's gaze in terms of concentration on the task and eye contact with the robot. We are also investigating the use of the facial node tracking software to visually recognize the speaker during a spoken interaction, raising the possibility of having the robot participate in turn-taking games between three or more players.

In terms of teaching with play, future efforts will be directed towards enhancing the "do as I do" imitation game to use visual sensing of body pose and motion. It would be particularly useful to be able to use such a technique to detect posture, as that is both a keen measure of the human's engagement and a method the robot could use to communicate its own interest in the game. We would further like to incorporate playful instruction into teaching Leo basic skills wherever it is practical to do so.

Finally, we will be using the advances in goal representation and sensing ability to create more complex and subtle games that incorporate more knowledge about the human's own mental state and goals. For example, Leo does not yet know precisely what the human's goal is if it is in opposition to his own. Learning such divergent collaborative aims will be an interesting task that has not been a principal focus of intelligent robotics work in the past. Ultimately, we hope to not just create robots that entertain us, but robots that know how to entertain us.

7. ACKNOWLEDGEMENTS

The work presented in this paper is a result of the ongoing efforts of the graduate and undergraduate students of the MIT Media Lab Robotic Life Group and our collaborators. Leonardo's speech understanding abilities are developed in collaboration with Alan Schultz and his group at the Naval Research Laboratory. Stan Winston Studio provided the physical Leonardo robot. Leonardo's cognitive architecture is built on top of the `C5M` code base of the Synthetic Characters Group at the MIT Media Lab. The Axiom ffT facial tracking software is provided as part of a collaboration with Nevengeneering, Inc. This work is funded in part by a DARPA MARS grant and in part by the *Digital Life* and *Things That Think* consortia.

8. ADDITIONAL AUTHORS

Additional authors: Andrea Lockerd, Hans Lee, and Cynthia Breazeal (Robotic Life Group, MIT Media Lab, email: {`alockerd, hcl337, cynthiab`}`@media.mit.edu`).

9. REFERENCES

[1] Baerends-Van Roon, J.M. and Baerends, G.P. *The Morphogenesis of the Behavior of the Domestic Cat, with a Special Emphasis on the Development of Prey-Catching.* North-Holland Publishing Co., Amsterdam; New York, 1979.

[2] Blumberg, B., Downie, M., Ivanov, Y., Berlin, M., Johnson, M.P., and Tomlinson, W. Integrated learning for interactive synthetic characters. In *Proc. 29th Annual Conf. on Computer Graphics and Interactive*

Techniques (SIGGRAPH '02), pages 417–426, San Antonio, Texas, 2002.

[3] Breazeal, C. *Sociable Machines: Expressive Social Exchange Between Humans and Robots.* PhD thesis, Massachusetts Institute of Technology, 2000.

[4] Breazeal, C. Towards sociable robots. *Robotics and Autonomous Systems*, 42(3–4):167–175, 2003.

[5] Breazeal, C., Brooks, A.G., Gray, J., Hancher, M., McBean, J., Stiehl, W.D., and Strickon, J. Interactive robot theatre. In *Proc. International Conference on Intelligent Robots and Systems*, Las Vegas, Nevada, October 2003.

[6] Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., and Blumberg, B. Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artificial Life*, 2004 (to appear).

[7] Call, J. and Carpenter, M. Three sources of information in social learning. In Dautenhahn, K. and Nehaniv, C., editors, *Imitation in Animals and Artifacts.* MIT Press, 2002.

[8] Carnegie Mellon University Robot Soccer. CM-RMP: Robot soccer Segway RMPs. http://www.cs.cmu.edu/ robosoccer/segway/.

[9] Darrell, T., Gordon, G., Woodfill, J., and Harville, M. Integrated person tracking using stereo, color, and pattern detection. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, June 1998.

[10] Davies, M. and Stone, T. *Mental Simulation.* Blackwell Publishers, Oxford, UK, 1995.

[11] Eveland, C., Konolige, K., and Bolles, R.C. Background modeling for segmentation of video-rate stereo sequences. In *Proc. Computer Vision and Pattern Recognition*, pages 266–272, 1998.

[12] Fernandez, R. and Picard, R. Signal processing for recognition of human frustration. In *Proc. IEEE ICASSP '98*, Seattle, WA, 1998. IEEE.

[13] Gordon, R. Folk psychology as simulation. *Mind and Language*, 1:158–171, 1986.

[14] Heal, J. *Understanding Other Minds from the Inside*, pages 28–44. Cambridge University Press, Cambridge, UK, 2003.

[15] Ivanov, Y.A., Bobick, A.F., and Liu, J. Fast lighting independent background subtraction. *International Journal of Computer Vision*, 37(2):199–207, June 2000.

[16] Kidd, C. Sociable robots: The role of presence and task in human-robot interaction. Master's thesis, MIT Media Lab, Cambridge, Massachusetts, June 2003.

[17] Maes, P., Darrell, T., Blumberg, B., and Pentland, A. The ALIVE system: Full-body interaction with autonomous agents. In *Proc. Computer Animation '95 Conference*, Geneva, Switzerland, April 1995.

[18] Mauldin, M. Chatterbots, TinyMUDs, and the Turing test: Entering the Loebner Prize competition. In *Proc. Twelfth National Conference on Artificial Intelligences (AAAI-94)*, Seattle, Washington, August 1994.

[19] Meltzoff, A. The human infant as imitative generalist: A 20-year progress report on infant imitation with implications for comparative psychology. In Galef, B. and Heyes, C., editors, *Social Learning in Animals: The Roots of Culture*, pages 347–370. Academic Press, New York, 1996.

[20] Meltzoff, A. and Decety, J. What imitation tells us about social cognition: A rapprochement between developmental psychology and cognitive neuroscience. *Phil. Trans. R. Soc. London B*, 358:491–500, 2003.

[21] Meltzoff, A. and Gopnik, A. The role of imitation in understanding persons and developing a theory of mind. In Baron-Cohen, S., Tager-Flusberg, H., and Cohen, D.J., editors, *Understanding Other Minds: Perspectives from Autism*, pages 335–366. Oxford University Press, Oxford, UK, 1993.

[22] Meltzoff, A. and Moore, M.K. Explaining facial imitation: A theoretical model. *Early Development and Parenting*, 6:179–192, 1997.

[23] Mota, S. and Picard, R. Automated posture analysis for detecting learner's interest level. In *Proc. IEEE Workshop on Computer Vision and Pattern Recognition for Human Computer Interaction (CVPRHCI)*, Madison, Wisconsin, June 2003.

[24] Niederberger, C. and Gross, M.H. Towards a game agent. Technical Report 377, Institute for Scientific Computing, ETH Zurich, September 2002.

[25] Perzanowski, D., Schultz, A.C., Adams, W., Marsh, E., and Bugajska, M. Building a multimodal human-robot interface. *IEEE Intelligent Systems*, 16(1):16–21, 2001.

[26] Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

[27] Reeves, B. and Nass, C. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge University Press, Cambridge, England, 1996.

[28] Terzopoulos, D., Tu, X., and Grzeszczuk, R. Artificial fishes with autonomous locomotion, perception, behavior, and learning in a simulated physical world. In *Artificial Life IV: Proc. Fourth International Workshop on the Synthesis and Simulation of Living Systems*, pages 17–27 (plus 3 color plates), Cambridge, Massachusetts, July 1994.

[29] Viola, P. and Jones, M. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 511–518, Kauai, HI, December 2001.

[30] Wood, D., Bruner, J.S., and Ross, G. The role of tutoring in problem-solving. *Journal of Child Psychology and Psychiatry*, 17:89–100, 1976.