Ensemble:
Fluency and Embodiment for Robots Acting with Humans

by

Guy Hoffman

M.Sc., Tel Aviv University (2000)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2007

Author__________________________
Program in Media Arts and Sciences
August 20, 2007

Certified by____________________
Cynthia Breazeal
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by____
Deb Roy
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Ensemble:
Fluency and Embodiment for Robots Acting with Humans

by

Guy Hoffman

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 20, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

Abstract

This thesis is concerned with the notion of *fluency* in human-robot interaction (HRI), exploring cognitive mechanisms for robotic agents that would enable them to overcome the stop-and-go rigidity present in much of HRI to date. We define fluency as the ethereal yet manifest quality existent when two agents perform together at high level of coordination and adaptation, in particular when they are well-accustomed to the task and to each other. Based on mounting psychological and neurological evidence, we argue that one of the keys to this goal is the adaptation of an embodied approach to robot cognition. We show how central ideas from this psychological school are applicable to robot cognition and present a cognitive architecture making use of perceptual symbols, simulation, and perception-action networks. In addition, we demonstrate that anticipation of perceptual input, and in particular of the actions of others, are an important ingredient of fluent joint action.

To that end, we show results from an experiment studying the effects of anticipatory action on fluency and teamwork, and use these results to suggest benchmark metrics for fluency. We also show the relationship between anticipatory action and a simulator approach to perception, through a comparative human subject study of an implemented cognitive architecture on the robot AUR, a robotic desk lamp, designed for this thesis.

A result of this work is modeling the effect of *practice* on human-robot joint action, arguing that mechanisms that govern the passage of cognitive capabilities from a deliberate yet slower system to a faster, sub-intentional, and more rigid one, are crucial to fluent joint action in well-rehearsed ensembles.

Theatrical acting theory serves as an inspiration for this work, as we argue that lessons from acting method can be applied to human-robot interaction.

Thesis Supervisor: Cynthia Breazeal
Title: Associate Professor of Media Arts and Sciences, Program in Media Arts and Sciences

Ensemble:

Fluency and Embodiment for Robots Acting with Humans

by

Guy Hoffman

The following people served as readers for this thesis:

Thesis Reader____

Alex Pentland
Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Thesis Reader____

Lawrence Barsalou
Professor of Psychology
Emory University

Of the same order as the impossibility of rediscovering an absolute level of the real, is the impossibility of staging an illusion.

Jean Baudrillard (1929–2007),
Simulacra and Simulations

Acknowledgments

A doctoral dissertation, while — to a certain extent — a solitary undertaking, is difficult to endure and impossible to complete without the help and support of many.

Thanks are due to my advisor, Cynthia Breazeal, who accepted me into her lab even though I displayed an obviously erratic personality both in my resumé and by applying once, declining, and then re-applying once more. She has since repeatedly proven her open-mindedness to my idiosyncrasies, and to the unusual type of student that is attracted to human-robot interaction. Whenever I walked into her office with that well-known feeling of doctoral despair, Cynthia has been able to pick up the randomly colliding ideas in my head and collect them into a coherent vision, argument, and game plan. I also consider myself lucky to have an advisor who invariably sides with her students vis-a-vis any outside forces, who makes a point of giving them ample opportunity to present their work and meet the leaders in our and other fields, and who trusts them to a courageous extent. Any graduate student, past or preset, can appreciate these privileges.

My gratitude also goes out to my thesis readers. Larry Barsalou's oeuvre has been a major inspiration for much of the work described herein, and having him agree to be on my committee proved to be not only an honor, but also a pleasure and a rescue ring. His answers to my questions were always extremely thoughtful, elaborate, and informed. Sandy Pentland has been a valuable veteran guide on this trek. Having someone on my committee who understood artificial perception better than most, dating back to the days when I was still learning calculus halfway around the globe, helped me at key points to realize which trails to pave and which to abandon.

Then, of course, my families. My mother has been patiently watching her son not only leave the home ground to faraway "America", but — perhaps even more painfully — accepted his decision to go back to school way past a comfortable settling-down age. My brother Dani and my sister Thalia lovingly supported me even when it meant to lend me to a foreign land and accepting my lack of communication when I was stressing about something that I'm sure did not make much sense to either of them. Thanks also to my extended family scattered around the world, and in particular to the Wieners, who were a familiar anchor in this strange country. I enjoyed the rare gift of having my cousin Daniela live a few blocks from my house for a year, enjoying her pampering and cooking skills.

To Talia, who taught me what love really means. Among others it apparently means unabashedly saying thing that read straight out of a bad 80s song. Her support was both concrete and overarching. It was through her eyes, her encouragement, and her example, that I could — for the first time in my life — follow my vision and walk down a path that, without her, I would have left for the safety of a compromise long ago.

My other family: my friends, the closest of which have put up with my behavior for over 15 years now. I particularly want to acknowledge Udi and Ori, who started out — as so many friends do — as strangers randomly tossed into one's environs in their late teens, but who end up shaping who you are more than anyone else in your life. Thanks for keeping my feet on the ground, my head in my dreams, and my hands in and out of trouble, whichever was appropriate at the time. Very special thanks go out to Oren and Orit, who shared these four years with me most closely, and who I can safely describe as my surrogate home away from home. Your care has rescued me again and again. Thanks also to the many friends who have been there for me, physically and electronically: Dror, Yair, Yariv, Moran, Dana, Yoad, Shahar, Josh, Alyssa, Keren, Moran, Shelly, Jeff, Tal, Galit, Neri, Maya, Nadav, Chris, Paula, Jeana, Nick, and the many others that my fried brain must have left out.

I once wrote that MIT, the Media Lab, and the Robotic Life group were three of the most impressive cultures I have ever been lucky enough to consider myself part of. I still hold this to be true. Anyone who has attended will attest that MIT is a truly unique place. Never before have I been surrounded by so large a group of people who are unrealistically brilliant, yet at the same time perfectly humble and generous. My lab mates at `robots`: Dan, Cory, Zoz, Matt, Mikey, Joon, John, and Fardad have been inspiring in their creativity and diligence, and unrelenting in their encouragement and support. My UROP Raymond has been an enormous help. And thanks to Polly, who has consistently been looking out for me. But I want to particularly point out the utter generosity of my co-pilot Jesse of RL-BP/1, and of Jeff and Andrea. Your almost daily help was as patient as it was indispensable.

Being at the Media Lab was a trip in and of itself. Having the daily privilege to roam the labs, offices, and spaces of this colorful cube, and being allowed to learn and play in what might well be the most satisfying playground for grown-ups, makes the incessant question "what can you do with a degree in Media Arts and Sciences?" laughably irrelevant.

Thanks also to my roommates at 45 Banks Street, past and present: Tristan, Maura, Sam, Mary, Carla, Edwina, Elizabeth, and the rest. Whatever other ups and downs I had, I never wanted to have another living arrangement. Thanks for allowing me to come home, and not only "go to my apartment", every night. And thanks to Dr. Jennifer Meyer, my therapist, who — literally — kept me sane over the last year.

Boston is not a friendly or particularly interesting urban center. Everyone seems to be too smart or rich for their own (and everyone else's) good. I'm therefore thankful for the proximity of the city of New York with its outgoing inhabitants. Your streets were a welcome refuge on many a weekend, and your wireless-enabled coffee shops harbors of semi-serious attempts at work. Thanks also to the 1369 coffee shop in Cambridge, and the Diesel Cafe in Somerville. The fact that I didn't want to pay for your criminally expensive wireless network facilitated many a long writing session without the distractions of modern communication.

Finally, I dedicate this work also to my father, who passed away nine years ago. His short life was not without tragedy, external and internal, one of the saddest of which was his constant feeling that he was not where he wanted to be, and wasn't doing what he wanted to do. This work, and the way I pursued it, is the beginning of my private attempt to correct this tendency in my own life.

Contents

List of Figures

List of Tables

Part I

Groundwork

Chapter 1

Introduction

Humans performing an activity together naturally converge on an impressive level of coordination, flexibility, and adaptation. Their timing is precise and efficient, they alter their plans and actions appropriately and dynamically, and this behavior emerges often without exchanging much verbal information.

We can marvel at the this sense of flow when we observe a well-rehearsed ensemble of performers or highly trained athletes, but in more subtle ways this behavior is apparent in almost any human interaction, be it in assembly tasks, game play, or dialog.

For lack of a better term, we denote this ethereal yet manifest quality of human interaction *fluency of joint action*, and in this thesis we will attempt to lay the groundwork enabling robots to perform more fluently with their human counterparts.

To date, human-robot interaction (HRI) is — more often than not — unintuitive, restrictive, and limited to a rigid command-and-response structure. Interaction with robots holds little of the fluent quality which is part of a satisfying collaboration, and this thesis puts forth that — true to the notion that interaction is rhythm as much as it is content[1] — robots must steer away from this trend and display a significantly more fluent meshing of their actions with ours, if they are truly to enter our daily lives in a socially meaningful manner.

[1] See Michalowski et al. [2007] for a review.

We believe that a promising route to achieve fluency in human-robot joint action[2] lies in subscribing to an embodied view of agent cognition. We argue that a core shortfall of existing systems is their inherent separation between perception, cognition, and action, whereas it is becoming increasingly clear that human perception and action are not mere input and output channels to an abstract symbol processor or rule-generating engine, but that instead thought, memory, concepts, and language are inherently grounded in our physical presence [Pfeifer and Bongard, 2007, Lakoff and Johnson, 1999, Barsalou, 1999, Wilson, 2001, 2002, Pecher and Zwaan, 2005].

Mounting evidence in psychology and neuroscience, as well as traditional knowledge in theater acting theory, indicates that our perceptual and motor systems shape the way we think and behave, and are intertwined more than originally assumed, forcing us to rethink our views of the human mind.

Recently, psychologists are also beginning to assess the neuro-cognitive mechanisms of social interaction and in particular those of joint action, investigating what enables us to coordinate our actions, and respond to each other with speed and precision [Barsalou et al., 2003, Sebanz et al., 2006]. Unsurprisingly, many of the ingredients underlying joint action rest in embodied aspects of the team members' cognitive processes, such as motor-based simulation and prediction, mirror neural systems, and top-down perceptual anticipation.

While part of the robotics research community has embraced an embodied view of artificial intelligence in the last two decades [Brooks, 1991, Pfeifer and Bongard, 2007], few if any of these lessons have been transferred from the realm of robotic cognitive modeling to the social aspects of HRI. This thesis hopes to begin to address this lack, investigating in particular the notions of perceptual concepts and simulation, as well as those of anticipation and emulation in human-robot interaction.

Anticipation is a concept that receives particular attention in this work, as we believe it to be crucial mechanism in achieving fluency in joint action. Anticipating world states, as well as the actions of a collaboration partner, enable an agent to time its actions pre-

[2] As well as a root motivation for trying to achieve that goal.

cisely[3] and — as we will show — has a significant effect on the human teammate's notion of the agent's commitment and contribution to the task. Anticipation is also related to a perception-centric view of cognition, through the proposed mechanism of emulation [Wilson and Knoblich, 2005], as well as to the motor-based and top-down view of perception advocated in this work [Wilson, 2001, Kosslyn, 1995].

That said, it is noteworthy that — while evident in isolated collaborations — fluency in joint action benefits greatly from repetition, practice, and rehearsal. A repetitive joint task becomes increasingly fluent, and some of the most rewarding and sought-after human performances (such as sports, martial arts, and performance arts) rely not only on talent, but on a rigorous training schedule tightly coupled with the principles of embodied cognition and repetitive action. In this thesis we investigate the relationship between fluency and practice and hope to address some aspects of practice in the context of perception and action, as well as in the context of simulation and anticipation. We believe that mechanisms that govern the passage of cognitive capabilities from a deliberate and flexible, yet comparatively slower system to a faster, sub-intentional, and more rigid one, are crucial to fluent joint action in well-rehearsed teams and ensembles.

A final note on theater acting: theories and methods of dramatic acting have inspired this work, and a chapter of this thesis is dedicated to HRI lessons learned from the Stanislavski system of acting, also known as "method" acting. Moreover, a robotic stage actor is an intriguing platform to demonstrate and evaluate the concepts and architectures described in this thesis, as acting is a highly embodied joint action, in which timing and fluency are crucial, and rehearsal is key to generate a tightly-meshed, individually-adaptive, and convincing performance. As very few examples of robotic stage actors are known to date, we have begun to implement and test the groundwork for a robotic stage actor. We hope, in future work, to apply more of the theories and systems described herein in a human-robot joint theatrical performance.

[3] See for example Weinberg and Driscoll [2006].

1.1 HRI Fluency: A review

Robots acting jointly with humans can be classified according to their degree of autonomy, ranging from full autonomy, through goal-based (mixed initiative) operation, via waypoint and heuristic specification, and down to full teleoperation [Goodrich et al., 2001]. This work is specifically concerned with an architecture supporting a fully or nearly autonomous robot, rather than with another area of research also coined "human-robot collaboration", but which is concerned with questions of control of a remote or local machine. Also, while some of the conclusions from this work could well be incorporated into this kind of robotic teleoperation research, such as that of Fong et al. [2003] or Jones and Rock [2002], these applications are not our primary concern.

Similarly, some literature frames human-robot collaboration in the context of mixed-initiative control and shared autonomy, arbitrating between the robot's autonomy and direct human control. These approaches also fail to address the question of shared-location fluency as we mean it in this thesis [Bruemmer et al., 2002, Goodrich et al., 2001].

This work, on the other hand, is mainly concerned with the emergence of fluent interaction between a human and a robot at a *shared location*, making use of the co-located partners' nonverbal behavior to achieve joint action fluency.

Within this definition, the concept of fluency in human-robot interaction has rarely been directly addressed, and only recently has begun to receive some marginal attention, such as in the works of Weinberg and Driscoll [2006], who reflect on nonverbal behavior and physically-based anticipation in their "Haile" robotic drummer project, or Michalowski et al. [2007], who study the rhythmic qualities of a beat-tracking dancing robot .

In contrast, the typical interactive robotic systems to date restrict their joint action with a human counterpart to a rather slow and heavily turn-based framework, making the pursuit of fluency a worthwhile endeavor.

For example, HERMES is a recent humanoid service robot for office environments [Bischoff and Graefe, 2004]. Its interaction with humans suffers from very structured turn-taking requirements including many pauses. Nonverbal behavior is practically absent on both the perceptual and the generative side. Other systems, even when they take into account the importance of a nonverbal channel of interaction (such as Sidner et al. [2003, 2006]), the dialog between the human and the robot is piecewise and feels more delayed than fluent.

The systems described are also mainly concerned with human-robot *dialog*, and many similar abound (and are usually afflicted by comparable deficiencies), e.g. Rybski et al. [2007], Argall et al. [2007], Doshi and Roy [2007], and Martinson and Brock [2007], just to cite from a recent conference on human-robot interaction . Robots that *act* jointly with humans, however, are few and far between.

Shared-location collaborative robots often include a robotic arm sitting vis-a-vis a human solving a task, such as the work of Kimura et al. [1999]. Their work addresses issues of vision and task representation, but does not investigate joint adaptation, and does not address the timing issue. In another case, a robot assists a human in an assembly task, and intervenes in one of three ways: taking over for the human, disambiguating a situation, or executing an action simultaneously with a human. This happens when the human asks for assistance or seems in need of help based on perceptual input [Sakita et al., 2004]. While relying on some nonverbal symbols, the result is still an interaction which is slow, heavily delayed, and stepwise, and contains nothing of the fluent quality we are striving for.

Some work in shared-location human-robot collaboration has been concerned with the mechanical coordination of robots in shared tasks with humans (e.g. Woern and Laengle [2000]). This work is predominantly concerned with single-action control and safety issues.

We have previously presented work in shared-location human-robot teamwork, investigating the role of nonverbal behavior on teamwork [Hoffman and Breazeal, 2004, Breazeal et al., 2005]. While this task-level decision system included turn-taking and joint plans, anticipatory action and fluency have not been addressed.

More recently, humans and robots have interacted in a soccer game setting, a natural and extremely promising testbed for human-robot fluency. However, human-robot join soccer games are at a stage of extremely initial efforts, in which the robot is given very little autonomy compared to the human team member [Argall et al., 2006].

Timing and synchronization have been reviewed on the motor level in the context of a human-robot synchronized tapping problem [Komatsu and Miyake, 2004]. Anticipatory action, without relation to a human collaborator, has been investigated in robot navigation work, e.g. Endo [2005].

In the field of robots for theater performance, most work has dealt with fully automated or extremely simple behavior on one end of the spectrum (for a detailed review, see Dixon [2004]) or fully teleoperated robots, such as the recent production of "Heddatron" [Les Freres Corbusier, 2006] on the other. Throughout the gamut, the focus of the work is usually on the *conceptual* relationship between man and machine, rather than the *interaction* between an autonomous robot and a human. In other work, robots have been pitted against each other on stage without the inclusion of a human scene partner [Bruce et al., 2000], but it is safe to say that fluent theatrical dialog between an autonomous robot and a human scene partner is still an unattained aim.

1.2 Contribution and structure of this document

This thesis proposes a new way to think about human-robot interaction by suggesting to examine the quality of interaction through the investigation of fluency in human-robot joint action. We propose an embodied view of joint action, and describe an implemented perceptual symbol architecture supporting this view, drawing conclusions of potential value for robot design, as well as for the endeavor of understanding the human mind when it acts in unison with another body. We support our designs with two human subject studies evaluating the contribution of perceptual symbols, anticipation, and simulation on human-robot fluency.

The remainder of Part I is dedicated to a survey of the theory that informed our work. In the next chapter, we document the shift in artificial intelligence and neuro-psychology towards an embodied view of cognition, and review theories of joint action.

Part II ("Fieldwork") sets out in Chapter 3 by discussing anticipation in joint action, describing a methodological framework and an experiment exploring the effects of anticipatory action on human-robot teamwork. This will be the first algorithmic foray into the notions of both fluency and practice. Chapter 4 introduces the perceptual research architecture developed as part of this work and shows our implementation of perception, simulation, and action. Chapter 5 integrates the previous two chapters and documents an implementation of the research architecture on a humanoid robot. This chapter relates anticipatory action to a perception-action based cognitive framework and points out new research questions arising from this effort, offering a more integrated view of the notion of practice. Chapter 6 describes a more mature and complex implementation of our approach on a non-anthropomorphic robot, leading to the human subject study concluding Part II in Chapter 7.

Part III ("Framework") commences in Chapter 8 by reflecting on three principles found throughout the acting method literature that may be of value to HRI design. Chapter 9 describes the physical design path of a robotic desk lamp, the non-anthropomorphic robot used in the experiments described in Chapter 7. Chapter 10 closes Part III discussing motivations to design non-humanoid expressive robots for HRI.

Finally, Part IV ("Closure") summarizes the core contributions of this thesis, and poses directions for extending the admittedly nascent work performed as part of it.

Chapter 2

Theoretical background

2.1 Embodied Cognition

During its first few decades of existence, artificial intelligence has not only *drawn* from theories of cognitive psychology, but to a wide extent also *shaped* notions of amodal, symbolic information processing. According to this view, information is translated from perceptual stimuli into nonperceptual symbols, later used for information retrieval, decision making, and action production. This view also corresponds to much of the work currently done in robotics, where sensory input is translated into semantic symbols, which are then operated upon for the production of motor control.

2.1.1 Perceptual symbol systems

An increasing body of recent findings challenges this view and suggests instead that concepts are grounded in modal representations utilizing many of the same mechanisms used during the perceptual process. A prominent theory explaining these findings is one of "simulators", siting memory and recall in the very neural modules that govern perception itself, subsequently used by ways of "simulation" or "imagery" [Barsalou, 1999, Kosslyn, 1995]. Perceptual symbols are organized in cross-modal networks of activation which are used to dynamically reconstruct and produce knowledge, concepts, and decision making. This view is supported by evidence of inter-modal behavioral influences, as well as by the detection of perceptual neural activation when a subject is using a certain concept in a non-perceptual manner (e.g. Martin [2001], Kreiman et al. [2000]).

Thus, when memory or language are invoked to produce behavior, the underlying perceptual processes elicit many of the same neural patterns and behaviors normally used to regulate perception [Spivey et al., 2005]. To name but a few examples, it has been shown that reading a sentence that has an implied orientation reduces response time on image recognition that is similarly oriented [Stanfield and Zwaan, 2001]; memory recall impairment is found to match speech impediments in children (mistaking rings for wings in children that pronounce 'r's as 'w's) [Locke and Kutz, 1975]; and comparing visually similar variations of a word (for example: a pony's mane and a horse's mane) is faster than visually distinct variations (e.g. a lion's mane) [Solomon and Barsalou, 2001].

2.1.2 Perception-action integration

In parallel to a perception-based theory of cognition lies an understanding that cognitive processes are equally interwoven with motor activity. Evidence in human developmental psychology shows that motor and cognitive development are not parallel but highly interdependent. For example, research showed that artificially enhancing 3-month old infants' grasping abilities (through the wearing of a sticky mitten), equated some of their cognitive capabilities to the level of older, already grasping[1], infants [Somerville et al., 2004].

A related case has been made with regard to hand signals, which are viewed by some as foremost instrumental to lexical lookup during language generation [Krauss et al., 1996], and is supported by findings of redundancy in head-movements [McClave, 2000] and facial expression [Chovil, 1992] during speech generation.

A large body of work points to an isomorphic representation between perception and action, leading to mutual and often involuntary influence between the two [Wilson, 2001].

[1] In the physical sense.

Some researchers speak of specific 'privileged loops' from perception to action, for example between speech and auditory perception, or visual perception and certain motor activity, indicating these action systems' roles in the perceptual pathway [McLeod and Posner, 1984].

2.1.3 Observation and anticipation

A number of experiments evaluating perceptual tasks have shown that task time is related to the mental simulation of kinematic configuration [Parsons, 1994]. In addition, neurological findings indicate that—in some primates—observing an action and performing it causes activation in the same cerebral areas, a phenomenon labeled "mirror neurons" [Gallese and Goldman, 1996]. Similarly, listening to speech has been shown to activate motor area related to speech production [Wilson et al., 2004], and in pianists, listening to a tonal sequence triggers neural activation in areas associate with finger movement — both for the current tone, and for the next tone in cases where the subject is familiar with the melody. This common coding is thought to play a role in imitation, and the relation of the behavior of others to our own, which is considered a central process in the development of a Theory of Mind [Meltzoff and Moore, 1997]. The predictive capabilities of these systems might also underlie the rapid and effortless adaptation to a partner that is needed to perform a joint task [Sebanz et al., 2006]. For a thorough review of these mirror neurological phenomena and the related connection between perception and action as it pertains to human-robot interaction, see Matarić [2002].

2.1.4 Top-down perception

Thus, perceptual processing should not be thought of as a strictly bottom-up analysis of raw available data, as it is often modeled in robotic systems. Instead, simulations of perceptual processes prime the acquisition of new perceptual data, motor knowledge is used in sensory parsing, and intentions, goals, and expectations all play a role in the ability to parse the world into meaningful objects. This seems to be particularly true for the parsing of human behavior in a goal-oriented and anticipatory manner, a vital component of joint action.

Experimental data supports this hypothesis, finding perception to be predictive (for a review, see Wilson and Knoblich [2005]). In vision, information is sent both upstream and downstream, and object priming triggers top-down processing, biasing lower-level mechanisms in sensitivity and criterion [Kosslyn, 1995]. Similarly, visual lip-reading affects the perception of auditory syllables indicating that the sound signal is not processed as a raw unknown piece of data [Massaro and Cohen, 1983][2]. High-level processing may also be involved in the perception of human figures from point light displays, enabling subjects to identify "complex actions, social dispositions, gender, and sign language" from very sparse visual information [Thornton et al., 1998].

2.1.5 Practice

Much of our critical cognitive capabilities occur in what is sometimes called a "cognitive unconscious" [Lakoff and Johnson, 1999], and it seems that these are more prominent in routine activity than those governed by executive control. Routine actions are faster and seem to operate in an unsupervised fashion, leading to a number of action lapses [Cooper and Shallice, 2000]. From introspection we know that supervisory control is often utilized when the direct pathways fail to achieve the expected results. These unsupervised capabilities seem to be highly subject to practice by repetition, and are central to the coordination of actions when working on a joint task, a fact that can be witnessed whenever we observe the performance of a highly trained sports team or a well-rehearsed performance ensemble.

2.1.6 Computational derivatives

The idea of top-down processing has been investigated and utilized in computational systems in the past, most notably in object and gesture recognition systems. Bregler demon-

[2] As cited in Wilson [2001].

strated a convincing application of this concept and applied it to action recognition in the visual field [Bregler, 1997]. Wren and Pentland created a robust human dynamic recognition and classification system by feeding likelihood data from high-level HMM procedures to pixel-level classifiers [Wren and Pentland, 1997, Wren et al., 2000]. Similarly, Hamdan et al. [1999] classified gesture sequences using Continuous Density Hidden Markov Models. Several collaborative planning systems integrated similar concepts in amodal settings (e.g. Rich et al. [2001]). However, the application of an integrated top-down implementation in an action-driven robotic behavior system is yet to be attempted.

Psychological evidence thus shows that concepts are not abstractions separated from their modal representation, but instead rely on the dynamic interaction of modus-specific activation. This occurs not only within, but also between modalities. In a computational system, this theory has received strong support through the work of Roy and Pentland [2002], showing a significant improvement in robustness of concept learning when multiple modalities were used.

This thesis will collectively use the term "embodied cognition" to relate to the effect and interrelation of the above mentioned elements: (a) perceptual symbol systems, (b) integration between perception and action, and (c) top-down processing. I use this overarching term to denote an approach that views mental processes not as amodal semantic symbol processors with perceptual inputs and motor outputs, but as integrated psycho-physical systems acting as indivisible wholes.

It is worthwhile to note that similar convictions have long been held by theorists and teachers of stage acting, who stress a psycho-physical unity underlying acting methodology. As early as the late 19^{th} century, François DelSarte noted that a "perfect reproduction of the outer manifestation of some passion, the giving of the outer sign, will cause a reflex feeling within."[Stebbins, 1887] With the rise of Stanislavskian method, this interrelation becomes even more prominent, and most teachers would probably agree with Boal in saying that "ideas, emotions and sensations are all indissolubly interwoven. A bodily movement 'is' a thought and a thought expresses itself in corporeal form."[Boal, 2002]

This thesis proposes to take a similar approach to designing a cognitive architecture for robots acting with human counterparts, be it teammates or scene partners. It is founded on the hope that grounding cognition in perception and action can hold a key to socially appropriate behavior in a robotic agent, as well as to context and temporally precise human-robot collaboration, enabling hitherto unattained fluidity in this setting.

2.2 Joint Action

Humans are exceptionally good at working in teams, ranging from the seemingly trivial (i.e. jointly moving a table through a doorway), to the complex (as in sports teams or corporations). What characteristics must a member of a team display to allow for shared activity? What rules govern the creation and maintenance of teamwork? And how does a group of teammates form individual intentions aimed to achieve a joint goal, resulting in a shared activity?

Joint action can best be described as doing something as a team where the participants share the same goal and a common plan of execution. The workings of this inherently social behavior have been of increasing interest to researchers in many fields over the past decade. Grosz [1996] — among others — has pointed out that collaborative plans do not reduce to the sum of the individual plans, but consist of an interplay of actions that can only be understood as part of the joint activity.

For example, if we were to move a table jointly through a doorway, your picking up one side of the table and starting to walk through the door does not make sense outside our joint activity. Even the sum of both our picking-up and moving actions would not amount to the shared activity without the existence of a collaborative plan that both of us are sharing (namely to move the table out the door), as well as the presence of tight action coordination before and during task execution.

The conceptual relationship between individual intentions and joint intentions is not straightforward, and several models have been proposed to explain how joint intentions relate to individual intentions and actions. Searle [1990] argues that collective intentions are not reducible to individual intentions of the agents involved, and that the individual acts exist solely in their role as part of the common goal.

In Bratman's detailed analysis of *Shared Cooperative Activity* (SCA), he defines certain prerequisites for an activity to be considered shared and cooperative [Bratman, 1992]. He stresses the importance of *mutual responsiveness, commitment to the joint activity* and *commitment to mutual support*. His work also introduces the idea of meshing singular sub-plans into a joint activity.

The Bratman prerequisites guarantee the robustness of the joint activity under changing conditions. In the table-moving example, mutual responsiveness ensures that our movements are synchronized; a commitment to the joint activity reassures each teammate that the other will not at some point drop their side; and a commitment to mutual support deals with possible breakdowns due to one teammate's inability to perform part of the plan. Bratman shows that activities that do not display all of these prerequisites cannot necessarily be viewed as teamwork.

Supporting Bratman's guidelines, Cohen and Levesque [1991] propose a formal approach to building artificial collaborative agents . Their notion of *joint intention* is viewed not only as a persistent commitment of the team to a shared goal, but also implies a commitment on part of all its members to a mutual belief about the state of the goal. Teammates are committed to inform the team when they reach the conclusion that a goal is achievable, impossible, or irrelevant. In our table-moving example, if one team member reaches the conclusion that the table will not fit through the doorway, it is an essential part of the implicit collaborative "contract" to have an intention to make this knowledge common. In a collaboration, agents can count on the commitment of other members, first to the goal and then—if necessary—to the mutual belief of the status of the goal. Some of these principles have been used in a number of human-robot teamwork architectures [Hoffman and Breazeal, 2004, Alami et al., 2005].

Joint Intention Theory predicts that an efficient and robust collaboration scheme in a changing environment requires an open channel of communication. Sharing information through communication acts is critical given that each teammate often has only partial knowledge relevant to solving the problem, different capabilities, and possibly diverging beliefs about the state of the task.

The above suggests that a central feature of any collaborative interaction is the establishment and maintenance of *common ground*, defined by Clark [1996] as "the sum of [...] mutual, common, or joint knowledge, beliefs, or suppositions". Common ground is necessary with respect to the objects of the task, the task state, and the internal states of the team members.

Common ground about a certain proposition p is believed by Clark to rely on a shared base b, that both suggests p and the common knowledge of b. People sharing a common ground must be mutually aware of this shared basis and assume that everyone else in the community is also aware of this basis. A shared basis can be thought of as a signal, both indicating the proposition p, and being mutually accessible to all agents involved. Clark coins this idea the *principle of justification*, indicating that members of a shared activity need to be specifically aware of what the basis for their common ground is [Clark, 1996].

This leads to the introduction of so-called *coordination devices* that serve as shared basis, establishing common ground. Coordination devices often include jointly salient events, such as gestural indications, obvious activities of one of the agents, or salient perceptual events (such as a loud scream, or a visibly flashing light).

Finally, an important type of reaching common ground inherent to teamwork is the principle of *joint closure*. Joint closure on a sub-activity is needed to advance a shared activity. Until joint closure is achieved, the sub-activity cannot be regarded as complete by a team. Joint closure is described as"participants in a joint action trying to establish the mutual belief that they have succeeded well enough for current purposes"[Clark, 1996].

To achieve joint closure, Clark argues that a *contribution*, i.e. a signal successfully understood, is usually made up of a presentation phase and a acceptance phase. For the contribution to be concluded, the presenter needs evidence of understanding on the interlocutor's part.

Both the principle of shared basis, and the principle of joint closure and justification call for the importance of mechanisms of *shared attention* between participants of a shared task. Shared attention is predominantly a nonverbal process. Thus, a robot working together with a human and following the principles of common ground must be able to understand nonverbal signals indicating a shared object of attention, as well as establish joint attention with the human through its own nonverbal behavior.

Another key capability of establishing fluent and correctly meshed joint action is the correct *anticipation* of a team member's actions. This anticipation occurs on several levels in human-human joint action. On the activity level, humans are biased to use an intention-based psychology to interpret other agent's actions [Dennett, 1987]. Moreover, it has been shown in a variety of experimental settings that from an early age we interpret intentions and actions based on *intended goals* rather than specific activities or motion trajectories [Woodward et al., 2001, Gleissner et al., 2000, Baldwin and Baird, 2001]. A person opening a door, once by using her right hand and once by using her left hand, is considered to have performed the same action, regardless of the different motion trajectories. But this person is doing a distinctly different action if she is opening a tap, even though the motion might be very similar to the one used for opening the door. To a human spectator, the goal of the action is often more salient than the physical attributes of it.

In teamwork, goals provide common ground for communication and interaction. Teammates' goals need to be understood and evaluated, and joint goals coordinated. When co-planning with a human team member, a robotic agent needs to take goals and their achievement into account when anticipating what the human's next action is probable to be.

On an action level, successful coordinated action has been linked to the formation of expectations of each partner's actions by the other [Flanagan and Johansson, 2003], and the subsequent acting on these expectations [Knoblich and Jordan, 2003]. In neural terms, motor resonance (the activation of motor areas when perceiving a similar motion from a conspecific) seems to be involved in motor prediction. A suggested mechanism for this is thought to be one of *emulators*, which can be thought of as equivalent to the simulators mentioned in the previous section, but operating one step ahead of current perceptual processing [Wilson and Knoblich, 2005]. These emulators may also play a role in the top-down perceptual processes discussed above.

Perhaps surprisingly, while the existence and complexity of joint action has been acknowledged for decades, the neuro-cognitive mechanisms underlying it have only received sparse attention, predominantly over the last few years (for a review, see Sebanz et al. [2006]). Much of the innovative work being done in the exploration of two humans sharing a task focuses on embodied structures of cognition, and recent work in neural imaging has repeatedly shown evidence for anticipatory structures which come into play during joint action between two humans [Sebanz et al., 2006]. This thesis proposes to adopt a similar approach for human-robot joint action.

Part II

Fieldwork

Chapter 3

Anticipation

In recent years, the cognitive mechanisms of joint action have received increasing attention [Sebanz et al., 2006]. Among other factors, successful coordinated action has been linked to the formation of expectations of each partner's actions by the other and the subsequent acting on these expectations [Knoblich and Jordan, 2003, Wilson and Knoblich, 2005]. We argue that the same holds for collaborative robots: if they are to go beyond stop-and-go interaction, agents must take into account not only past events and current perceived state, but also expectations of their human collaborators.

In this chapter we discuss the first pillar of our human-robot fluency program, an adaptive anticipatory action selection mechanism for a robotic teammate. We show a probabilistic approach, and analyze our model in a cost-based framework of coordinated shared-location action. We compare our approach to a purely reactive agent acting within a traditional perception-action loop, demonstrating a theoretical improvement in joint task efficiency.

We then present results from a study involving untrained human subjects working with a simulated version of a robot using our anticipatory system. We show a significant improvement in best-case task efficiency when compared to users working with a purely reactive agent. However, we were not able to show this difference being significant when measuring the mean score over repetitions. We attribute this to the small number of repetitions used in our study.

That said, we are not interested solely in efficiency, but also in the qualitative notion of fluency in coordinated action meshing, ultimately leading to more appropriate collaborative behavior. In a post-study survey we found a significant difference in the perceived contribution of the robot to the team's fluency and success, as well as its commitment to the team. Given that there are no generally accepted measures of teamwork fluency, we raise three fluency metric hypotheses, and evaluate these between the two conditions. We find the groups to differ significantly in two (time between human and robot action, and time spent in concurrent motion), but not in a third (human idle time).

3.1 World Description

In this chapter, we model the team fluency problem as a discrete time-based deterministic decision process including two agents, a *robot* and a *human*, working together on a shared task.

Both robot and human share a common workspace, which at any time point is in one of a finite number of states $\Sigma_W = \{s^w_0, \dots, s^w_n\}$, and is initially in state s^w_0. The agents also have a number of states $\Sigma_H = \{s^h_0, \dots, s^h_n\}$ (the human's states) and $\Sigma_R = \{s^r_0, \dots, s^r_n\}$ (the robot's states). In our model the robot can only perceive the state of the workspace if it is in a subset of states, called *perceptive* states. We shall denote a full state of the system $s_n \equiv < s^w_i, s^h_j, s^r_k >$, and similarly, $\Sigma = \Sigma_W \times \Sigma_H \times \Sigma_R$.

Human and robot have distinct abilities, described as two sets of actions, $A_H = \{a^h_1, \dots, a^h_k\}$ for the human, and $A_R = \{a^r_1, \dots, a^r_l\}$ for the robot.

$T : ((A_H \cup A_R) \times \Sigma) \mapsto \Sigma$ is a transition function that maps certain state-action pairs to new states. We denote a particular state-action pair transition in T

$$\tau^x_i(s_k) = s_l \equiv < a^x_i, s_k > \mapsto s_l$$

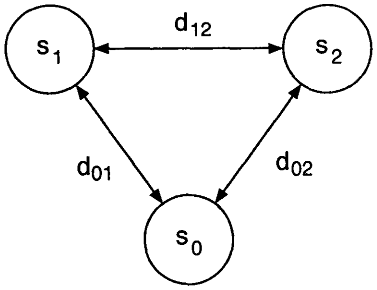


Figure 3-1: Transition costs between two states as a directed graph.

meaning that if agent $x \in \{h, r\}$ (human or robot) performed action i while the world is in state s_k, the world would transition into state s_l after applying the action.

A central motivation of our model is to investigate aspects of *time* associated with actions of two collaborating agents. Therefore, state transitions are not atomic, and the decision to take a particular action does not result in an immediate state transition. Instead, moving between states takes time, and is associated with a known discrete cost, which is a function of the states before and after the action. This cost can be thought of as the 'distance' between states, or more generally — the duration it takes to transition between states. We denote the cost of transitioning between states s_k and s_l with $d(s_k, s_l)$.

$$D : \left(\bigcup_{i \in W,H,R} < \Sigma_i \times \Sigma_i > \right) \mapsto \mathbb{N}$$

Thus when, at time t, agent $x \in \{h, r\}$ decides to take action a_i^x on state s_k and $\tau_i^x(s_k) = s_l$, the world will be in state s_l only at time $t + d(s_k, s_l)$. While the other agent may take more actions during this time, the next time step agent x will be able to take another action is $t + d(s_k, s_l)$. It can be useful to depict the state transitions as a directed graph, with the nodes representing the states and the edges the transitions between the states, weighted by the duration/cost function D (see Figure 3-1). For sake of simplicity, we will sometimes denote $d(s_k, s_l)$ as d_{kl} as indicated in the figure.

Agents cannot change the other agent's states with their actions, but they operate on a common workspace. Therefore, our model is clearly ill-defined with regards to race conditions on the Σ_W state space. There are several possible solutions (such as the use of semaphores and other synchronization mechanisms). In this work, for the sake of simplicity, we will assume that actions that change the workspace are locking with regards to actions that operate on the common workspace *for both agents*. In the implementation of our model described below, we solve this race condition by making all state transitions effecting the workspace atomic.

3.1.1 The Factory World

In our experiments we use a simulated factory setting (Figure 3-2). The goal of the team is to assemble a cart made of a *Body*, a *Floor*, two *Axle*s, and four *Wheel*s. The various parts have particular ways to be attached to each other — the *Body* is welded to the *Floor*, *Axle*s are riveted to the *Floor* and *Wheel*s are attached to *Axle*s using a wrench of matching color. A *component* is a partially assembled cart segment that includes one or more individual parts attached to each other, for example $Axle + Body + Floor$.

The labor is divided between the human and the robot: the human has access to the individual parts, and is capable of carrying them and positioning them on the workbench. The robot is responsible for fetching the correct tool and applying it to the currently pertinent component configuration in the workbench. Each part has a stock location (with an infinite supply of parts), and each tool has a storage location, to which it has to be returned for the robot to be able to find it again. The workbench can, at any one time, contain at most two components.

In the above-described framework, the workbench state space $\Sigma_W = Comps^2$, where $Comps$ is the space of all possible components.[1] In our case, $|Comps| = 42$, and thus $|\Sigma_W| = 42 \times 42 = 1764$. The robot's state space includes its position at one of the four tools' storage areas or at the workbench, and whether the robot is holding one of the tools or not. Therefore,

[1] Note that this is not 2^{Parts}, since not all parts can be attached to each other, and some parts can appear multiple times in a component.

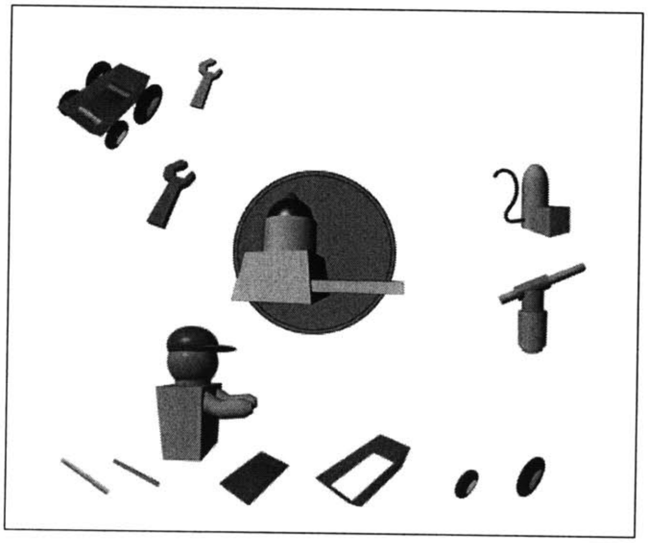


Figure 3-2: Simulated factory setting with a human and a robot building carts, while sharing a workbench (gray circle), but dividing their tasks. The robot has access to the tools (right and top-left of workbench), whereas the human is responsible to bring the parts (below the workbench). Top left shows a completed cart.

$|\Sigma_R| = 25$. Similarly, with 6 kinds of parts, $|\Sigma_H| = 49$. Thus, the size of the state-space in this simulation is 2,160,900.

The action-space of the robot is

$$\begin{aligned} A_R = \{&Workbench, Welder, Rivet, Wrench1, \\ &Wrench2, PickUp, PutDown, Use\} \end{aligned}$$

The first four actions are mobility actions, moving to one of the four locations in the factory. *PickUp* and *PutDown* are operational only in the tool locations, with the latter only available at the correct storage location of the currently held tool. A_H is a similar space with two more navigation action, and no *Use* action. To illustrate the state transitions, here are two examples of transitions brought about by actions in A_R (Here, W is *Wrench1*, R is

RivetGun):

$$\tau^r_{Pickup}(<s^w_i, s^h_j, <W, \emptyset>>) = <s^w_i, s^h_j, <W, W>>$$

$$\tau^r_{Use}(<<Floor, Body>, s^h_j, <Wrkspc, Welder>>) =$$
$$= <<Floor + Body, \emptyset>, s^h_j, <Wrkspc, Welder>>$$

$$\tau^r_{RivetGun}(<s^w_i, s^h_j, <W, \emptyset>>) = <s^w_i, s^h_j, <R, \emptyset>>$$

The duration cost of a state transition that involves navigation is the distance between the previous and the new location. The duration cost for state transitions involving the inventory of an agent, or changes to the workbench, is 1 in this implementation, but could theoretically be different for each tool.[2]

The robot can perceive the state of the workbench only when it is located in it. Workbench state changes that happen while the robot is in any other state are not applied to its internal representation.

Moreover, we assume that the robot has a function Φ that maps the workbench state to the appropriate tool required to bond the two components on the workbench. For example: $\phi(<Floor + Axle1, Wheel1>\ ) = Wrench1$. This can be a lookup table, or a decision process. In our implementation, the agents models the components as having open and closed male and female attachments to deduct Φ. Also note that some workspace states do not warrant any tool, because they either have an empty component, or two components that cannot be attached. We mark these function values as $\phi(s^w_i) = \emptyset$.

[2] For example: welding can take longer than riveting, and picking up the wrench could be faster than picking up the welder.

3.2 Reactive Agents

A baseline agent that is purely responsive to its environment and internal state, can be defined by an action policy that waits in the workbench when $\Phi(\mathit{WorkBench}) = \emptyset$, and fetches tool x, uses it, returns it, and returns to the workbench when $\Phi(\mathit{WorkBench}) = x$.

The obvious fallacy of this policy occurs when the same tool is needed twice in a row (which can happen with the wheels and axles, in the factory domain), resulting in a superfluous sequence of returning and then fetching the same tool. If the distance between the workspace and the tool is d, and under the assumption that the time it takes for the human to bring the next part h is smaller than $4d + 3$, the total cost of this sequence is $6d + 5$.

The naïve policy can therefore be improved by delaying the decision to return a tool until the state of the workbench changes. This prevents the agent from returning a tool before it is certain that it is not needed again in the next step. We call this policy *conservative tool return*. Given the time delay between the two *Uses*

$$\delta = \begin{cases} h - (2d + 2) & \text{if } h > 2d + 2, \\ 0 & \text{otherwise.} \end{cases}$$

The total cost of the sequence is $2d + 3 + \delta$. The gain in performance is $4d + 2 - \delta$.

However, it is straightforward to demonstrate that there is a negative impact of the "conservative tool return" strategy in the case where the next tool needed is different than the current tool. Note that the cost effect of conservative tool return is dependent not only on the known world configuration, but also on the turnaround time of the human action h, a quantity that can not be known but only estimated by the robotic agent. Additionally, the overall expected cost effect is dependent on the probability distribution on the workbench configuration over time. It therefore makes sense to discuss an action selection policy based on these factors, which is the topic of the following section. We will then frame the two reactive policies discussed here as a subset of the proposed anticipatory policy.

3.3 Anticipatory Action Selection

As discussed in the introduction, humans are remarkably adaptive and increasingly effective when performing repetitive trials of an identical task collaborating with a consistent teammate. The use of educated anticipatory action based on expectations of each other's behavior may be a key ingredient in the achievement of this action fluency. In this section we will attempt to adopt this insight in the human-robot interaction domain within the discussed framework.

A necessary assumption for anticipatory action selection in our agent is that the human collaborator will follow a roughly consistent action pattern, i.e. will make similar decisions under similar circumstances.

The agent thus models the workbench as a first-order Markov Process.[3] The probability of the workbench state at time t, σ_t^w, is thus conditional on σ_{t-l}^w and denoted as

$$p_{i|j}^w \equiv Pr(\sigma_t^w = s_i | \sigma_{t-l}^w = s_j)$$

The agent can learn the parameters of this Markov process using a naïve Bayesian estimate. To do this, the agent keeps a one-step history of the state transitions of the workbench. A change from state s_j to state s_i increases the counter $n_{i|j}$. Consequently, $p_{i|j}^w$ is computed as

$$p_{i|j}^w = \frac{n_{i|j}}{\sum_{k=1}^{|\Sigma w|} n_{k|j}}$$

However, in order to estimate the cost of preemptive action as described in the following section (which is ill-defined for non-constructive workbench states), and also to reduce the decision state space, the robot in our factory domain can alternatively model the proba-

[3] A presumably more realistic model would be to view the collaboration as a Hidden Markov Model, with the human state transitions being hidden, and the workbench transitions being the evidence layer of the model. However, since many of the human's state transitions do not affect the workbench state, and the probability of workbench transitions conditional on the human state transitions $Pr(\sigma_t^w = s_i | \sigma_t^h = s_j)$ are not independent of σ_{t-l}^w, it is unclear whether such a model would indeed be of value in our domain, and is therefore left to future investigation.

bility of the tool needed based on the previous state: if $Q(x) = \{s_i : \phi(s_i) = x\}$ is the set of workbench states that warrant tool x, the new probability model learned is now

$$p_{x|j} \equiv Pr(\sigma^w_t \in Q(x)|\sigma^w_{t-l} = s_j)$$

We estimate this model as follows: a change from state s_j to state $s_i \in Q(x)$ increases the counter $n_{x|j}$. Using a Laplace correction of 1 [Kohavi et al., 1997], $p_{x|j}$ is then estimated by

$$p_{x|j} = \frac{n_{x|j} + 1}{\sum_{k=1}^{|Tools|} n_{k|j} + 1}$$

3.3.1 Action Selection

As the agent only perceives the workbench state (and therefore information about the transition distribution) when it is in the workbench state, it makes sense to make decisions in terms of *action sequences*. The acquisition of these sequences is beyond the scope of this work, but suffice to say that in our scenario the agent needs only to consider action sequences that begin and terminate while it is in the workbench state.

In the discussed factory domain we can identify four proto-sequences (state transitions for the full sequences are presented in square brackets):

1. Pick up a tool and use it
   $[< s^r_0, \emptyset > \rightarrow < s^r_0, x >]$
2. Return a tool and return to workbench
   $[< s^r_0, x > \rightarrow < s^r_0, \emptyset >]$
3. Return a tool, bring a new tool, and use it
   $[< s^r_0, x > \rightarrow < s^r_0, y >]$
4. Do nothing and wait
   $[s^r_i \rightarrow s^r_i]$

The action selection process operates as follows: at any time the robot is in the workbench state, it evaluates the cost of each of the proto-sequences. Proto-sequence 1 needs to be grounded for each tool and proto-sequence 3 needs to be grounded for each of the currently not held tools. Given the probability distribution, the robot can compute the expected cost for choosing each of the strategies, and selects a grounded sequence optimizing for cost. In calculating the expected cost for proto-sequences 1–3, we need to assume that $\forall i.[h < 2d_{0i}]$. Also note that the cost in our calculations includes performing the correct action afterwards. Denoting the current state of the workbench s_j, and the workbench position 0, the expected duration cost of proto-sequence 1–3 are

$$\begin{aligned} Cost_1(x) =& p_{x|j}(2d_{0x} + 2) + \\ & \sum_{y \neq x} [p_{y|j}(3d_{0x} + d_{xy} + d_{0y} + 4)] \end{aligned}$$

$$Cost_2(x) = \sum_{y=1}^{|Tools|} [p_{k|j}(2d_{0x} + 2d_{0y} + 3)]$$

$$\begin{aligned} Cost_3(x, y) =& p_{y|j}(d_{0x} + d_{xy} + d_{y0} + 3) + \\ & \sum_{z \neq y} [p_{z|j}(d_{0x} + d_{xy} + 2d_{y0} + d_{yz} + d_{z0} + 5)] \end{aligned}$$

Action sequence 4 is unique insofar as it is dependent not only on the state transitions in the workbench, but also on the behavior of the human teammate. If the human's next workbench-changing action is at time $t + h$, the cost of waiting is the cost of performing the correct action with complete confidence, plus h. For the case that the robot is holding a tool z:

$$Cost_4 = p_{z|j} + \sum_{y \neq z} [p_{y|j}(d_{0z} + d_{zy} + d_{y0} + 3)] + h$$

For the case that the robot is not holding a tool:

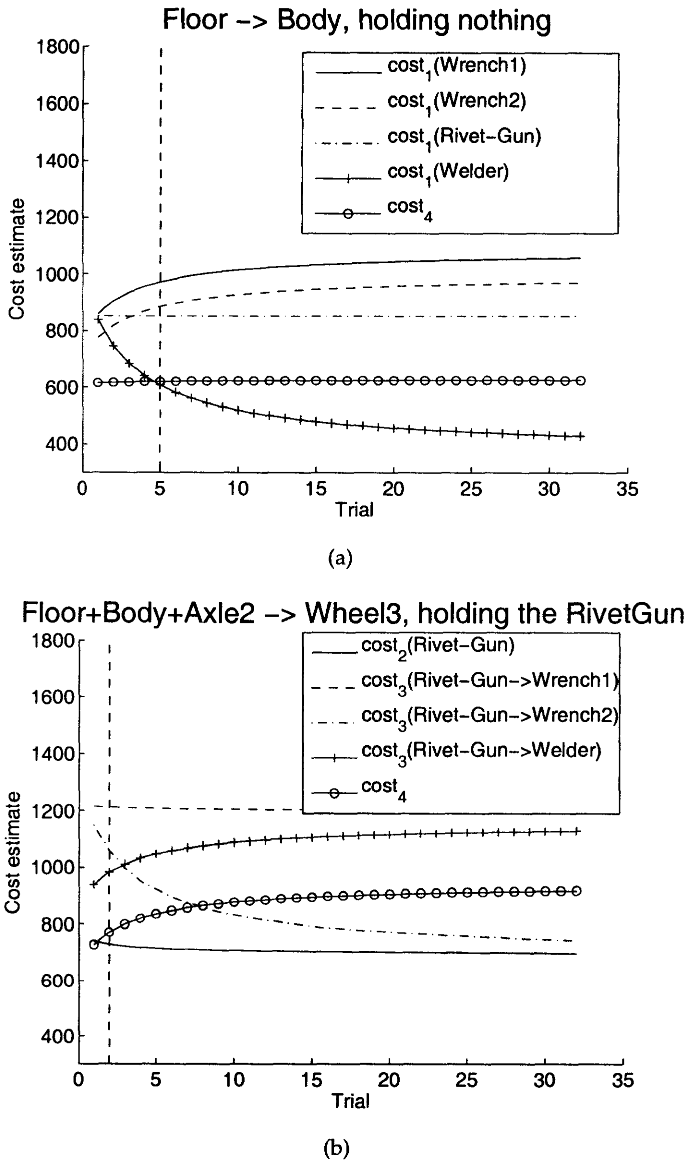
$$Cost_4 = \sum_{y=1}^{|Tools|} [p_{y|j}(2d_{0y} + 2)] + h$$

However, Since h is not directly accessible to the robotic agent, its estimate can be used as a confidence parameter, adjusting between an aggressively anticipatory behavior and a more cautious approach (see also below).

Using the above notation, we can now rephrase the previously discussed reactive agent behaviors. The naïve agent's policy can be viewed as selecting proto-sequence 2 whenever it is holding a tool in the workbench, and selecting proto-sequence 1 whenever a tool is warranted. The agent employing conservative tool return can be rephrased as selecting proto-sequence 4 whenever no tool is warranted, and selecting proto-sequence 1 or 3 if a workbench state warrants a tool. This rephrasing enables comparison between the different policies, as described in the following section.

3.3.2 Analysis

Figure 3-3 demonstrates the adaptation of cost-optimizing anticipatory action vis-a-vis a consistent human teammate with $h = 250$ and the factory layout as depicted in Figure 3-2. Figure 3-3(a) depicts the expected cost for the five available action sequences when the robot perceives the *Floor* in the workbench, holding nothing, over 31 trials in which the human consistently brings the *Body* in this situation. We can see that Sequence 4 (waiting) is the cost-optimizing action for trials 1–4, and that getting the *Welder* becomes the optimal anticipatory action from trial 5 onwards. In contrast, Figure 3-3(b) shows that when holding the *RivetGun* and perceiving *Floor* + *Body* + *Axle*2 (with a human consistently bringing *Wheel*3 to the workbench), Sequence 2—returning the *RivetGun*—becomes



(a)

(b)

Figure 3-3: Cost evaluation with a simulated human teammate. In (a) the robot is holding nothing and perceiving the *Floor* with a consistent human always bringing the *Body* next; in (b) the robot is holding the *RivetGun* and perceiving *Floor* + *Body* + *Axle2* with a consistent human bringing *Wheel3*.

optimal starting from the second trial. This difference becomes apparent considering the location of the *Wrench* on the opposite side of the workbench, making it considerably more expensive to wait, the more confident the robot gets that the *Wrench* is needed next. It it interesting to note that due to the particular tool arrangement, returning the *RivetGun* and pre-fetching the *Wrench* does not become cost-optimizing even after 31 trials. While it does become more optimal than waiting after eight consistent trials, the cost of an erroneous prediction, even as it becomes extremely unlikely, is still too high, resulting in a preference for Sequence 2 over Sequence 3. Note that this does not hold for other decision junctions. For example, holding the *Welder* with a consistent need for the *RivetGun*, prefetching it on the way back from the *Welder* location becomes cost-optimizing on the sixth trial (not shown in figure).

Using the analysis in the previous section, we can now compare the reactive agents to our proposed method. In the case described in Figure 3-3(a), our algorithm is equal to the reactive agents (equivalent to Sequence 4) in trials 1–4, and outperforms them increasingly as the amount of evidence increases. In the case described in Figure 3-3(b), the naïve reactive agent is equivalent to Sequence 2, slightly outperforming our method in the first trial, and then matching it, while the conservative tool return agent (equivalent to Sequence 4) chooses a more costly approach than our method from trial two onward. Generally speaking, using $h = 250$ in the factory scenario, we usually see the agent outperforming the reactive agents within 2 trials, and converging into full anticipatory behavior within 10 trials.

In actual trial runs with an experienced and consistent human teammate, we can see evidence to that effect. Whereas the reactive agent with conservative tool return remains constant at a construction cost[4] of circa 800, the anticipatory adaptive agent shows a significant improvement after the first trial and again at the sixth trial, finally settling at a lower per-cart construction cost of circa 650 (Figure 3-4).

Figure 3-5 shows the effects of inconsistency on the human teammate's part. Given the

[4] The cost units, when measured with a human teammate, are in simulation frames, running at 30 frames per second.

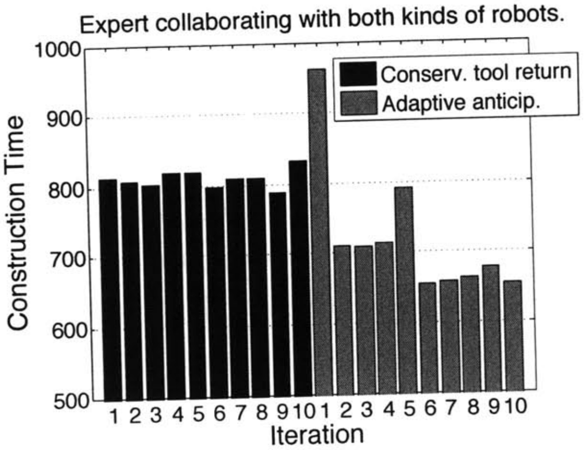


Figure 3-4: Change in per-cart construction time with an expert consistent human vis-a-vis the reactive agent (left) and the adaptive anticipatory agent (right).

Floor in the workbench, the human in this case brings the *Body* with a probability of 70%, and an *Axle* with a probability of 30%. The result is that waiting for the human's next move remains cost-optimizing for 12 iterations, delaying the anticipatory behavior of the agent and resulting in slower convergence into a fluent and efficient activity pattern.

A final note regarding the risk-taking parameter h, which we defined as the estimated time the human's turn takes: varying h (Figure 3-6) affects the relative optimality of Sequence 4 (waiting for the human). Lowering h significantly corresponds to an expectation that the human returns very quickly with the next part, resulting in a risk-averse policy. At the junction depicted in Figure 3-3(a), for example, lowering h to 10 will render the decision function equivalent to the 'conservative tool return' reactive agent discussed above. Setting $h = 500$ results in an agent performing anticipatory actions as soon as trial one. Fixing h at 100 results in taking the correct anticipatory action at trial 16, instead of trial 5.

Ideally, h should be specific per state, as well as learned over time as the agent collects more data regarding the turnaround time of the human teammate.

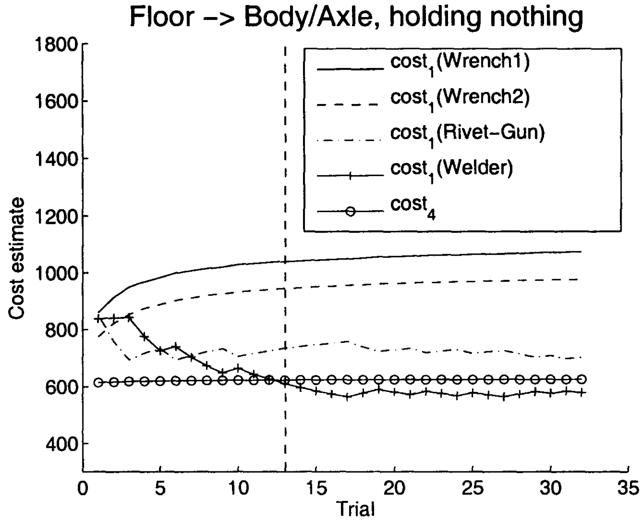


Figure 3-5: Effect of an inconsistent human teammate: graph shows perceiving the *Floor* with an human teammate bringing the *Body* with a probability of 70% and an *Axle* with a probability of 30%.

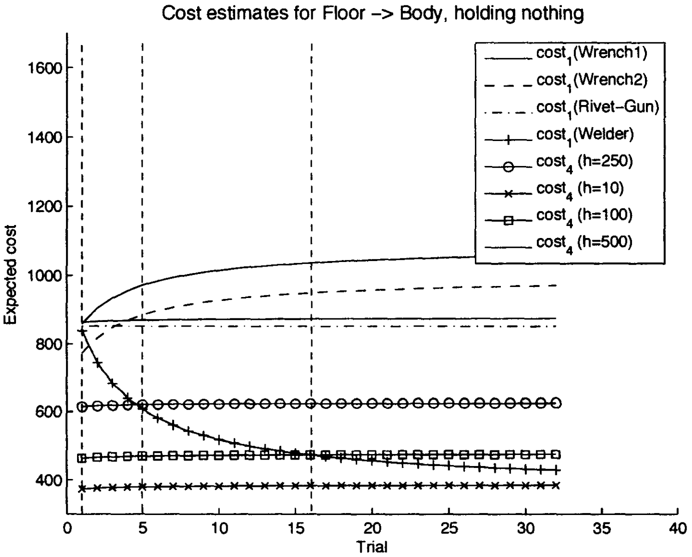


Figure 3-6: Cost analysis perceiving the *Floor* with a consistent human teammate, but varying the estimated human turnaround cost h.

3.4 Human Subject Study

To further investigate the effect of adaptive anticipatory action selection, we conducted a human subject study. We expected to see an increase in efficiency as predicted by the theoretical analysis, as well as an increase in the perceived contribution of the robot to the team's fluency and success.

3.4.1 Experimental Design

We recruited 32 participants (15 female) from the MIT community through email solicitation and posters. Participants arrived at our laboratory and were arbitrarily assigned to one of two experiment conditions. Subjects in *Group A* (the `REACTIVE` condition) interacted with a reactive agent using the "conservative tool return" policy; those in *Group B* (the `ANTICIPATORY` condition) interacted with an anticipatory agent.

All the participants received the following identical instructions (edited for brevity, omitting user interface instructions):

> In this study you play a video game. This game has two characters, Symon, a forklift-like robot in a cart factory and an avatar representing you, the human. Symon is surrounded by four tools: the welder, the rivet gun and two wrenches. The human is surrounded by six kinds of cart components: a floor, a body, two kinds of axles, and two kinds of wheels. In the center of the screen is a round workspace.
>
> In this game your goal is for the human-robot team to build 10 carts. Each of the team members has their own role in this joint effort. The human's role is to bring components to the workspace, the robot's role is to attach the car parts using the tools. The following tools attach the following components:
>
> 1. The wrench attaches a wheel to the matching color axle
> 2. The welder attaches the floor to the body

> 3. The rivet gun attaches both the axles to the floor
>
> A complete car has one floor, one body, two axles (one of each kind), and four wheels (two of each kind).
>
> The robot can only use a tool if there are exactly two cart parts in the workspace. Each of these parts can be made up of more than one simple components. For example - the workspace could contain one part made up of an axle with two wheels, and one part made up of a floor with a body attached to it. In this case the robot could use the rivet gun. If there are more or less than two parts in the workspace, the robot can't do anything.
>
> Your goal is to build cars in the least amount of time. A cart's construction time is measured from the moment the first part is dropped in the workspace and until the cart is completed. You can always see your best score and your last score, as well as the all-time best score, in the corner of the screen.

The instructions were phrased so as to imply the importance of the team as a joint performing entity. To control for instruction bias, neither group was told whether the robot will adapt to their behavior. Before beginning the experiment, participants were allowed to practice for an unlimited amount with the software, set to their assigned experimental condition.

3.4.2 Results

Of the participants, five had to be eliminated from the study. Two violated the experimental protocol, one experienced a software crash, one was significantly inattentive, resulting in scattered behavior, and for one subject the logging functionality was not working, resulting in a loss of data. This left us with 27 subjects, 14 in Group A and 13 in Group B. All 32 completed a post-study survey regarding their experience.

Table 3.1 shows total cart construction measures for the population. Cost units are in simulation frames at 30 frames per second.

Table 3.1: Total cart completion metrics for untrained human subjects in the reactive (Group A) and adaptive anticipatory condition (Group B). We compare each subject's best score in ten trials, mean score over ten trials, and tenth trial, using a T-test for independent samples

Score metric	REACTIVE mean	REACTIVE std.dev.	ANTICIPATORY mean	ANTICIPATORY std.dev	t(25)
Best	1091.6	200.5	930.1	105.6	2.59; $p < 0.02$
Mean	1423.5	328.6	1233.3	227.5	1.73; not signif.
Final	1182.4	274.3	1030.7	154.8	1.75; not signif.

Each subject's best performance is significantly better at a confidence level of 98% in the adaptive anticipatory case compared to the reactive case. Measuring the mean construction time over ten trials, as well as the time for construction of the tenth cart, we find the subjects in the anticipatory case to be better (at $p < 0.1$), but not significantly at a 95% confidence level. We believe that this is in part due to the fact that several subjects in Group B took a number of inconsistent trials to identify that the robot was adaptive, leading to a convergence to a stable construction pattern only in the last few carts (see also: Section 3.4.3). According to this hypothesis, both the mean and the final cart construction cost would be significantly lower in the anticipatory case if there were more trial runs per subject.

Survey

In the post-experimental survey, we found significant differences between participants in the two groups. On a seven-point Likert scale, subjects in the ANTICIPATORY condition selected a significantly higher mark than those in the REACTIVE condition, when asked whether "The robot's performance was an important contribution to the success of the team", "The robot contributed to the fluency of the interaction", and "It felt like the robot was committed to the success of the team". Figures 3-7 through 3-9 show the differences

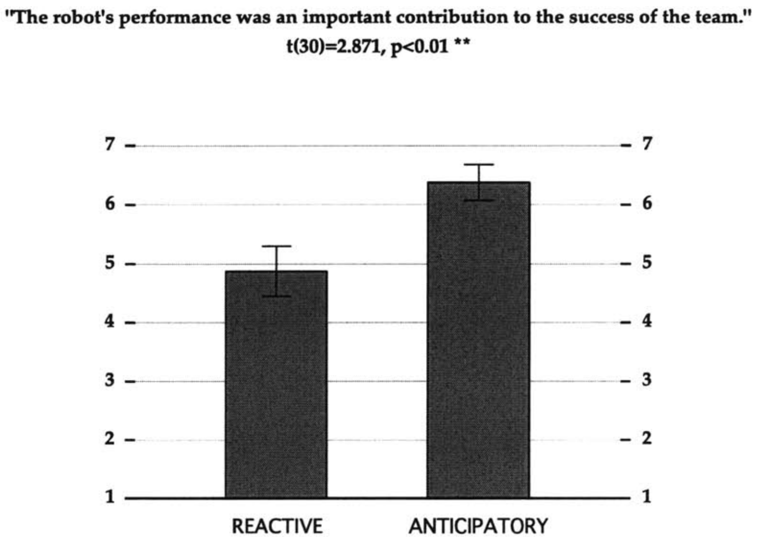


Figure 3-7: Significant self-report difference between participants in the *REACTIVE* (Group A) and the *ANTICIPATORY* (Group B) conditions, using a T-test for independent samples. Error bars indicate standard error.

in self-report between the two conditions.

The two groups did not differ significantly when subjects were asked whether they themselves were "committed to the success of the team", or whether they "trusted the robot to do the right thing at the right time." Both groups averaged between 6 and 7 on these two questions.

Measures of Fluency

In sum, we found significant differences between the two conditions in the subjects' *perception* of fluency as well as in their perception of the robot's commitment and contribution to the team's success. This conclusion is further embellished by the qualitative findings described in Section 3.4.3 below. At the same time, the mean (and convergent) task efficiency of the team was not significantly different between the conditions. This contradictory phenomenon could suggest that the notion of fluency, commitment, and appropriate

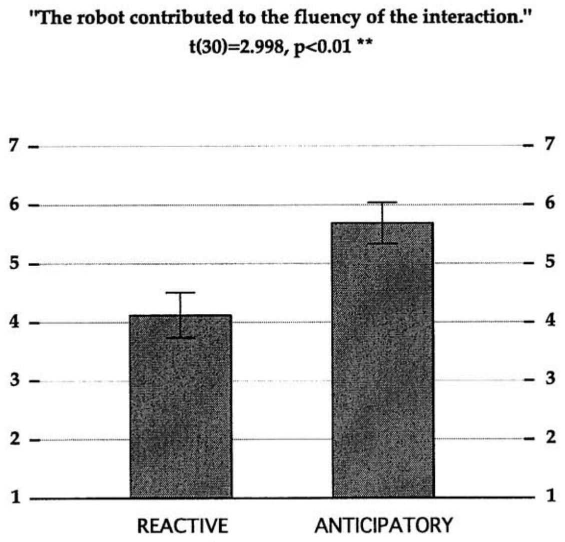


Figure 3-8: Significant self-report difference between participants in the *REACTIVE* (Group A) and the *ANTICIPATORY* (Group B) conditions, using a T-test for independent samples. Error bars indicate standard error.

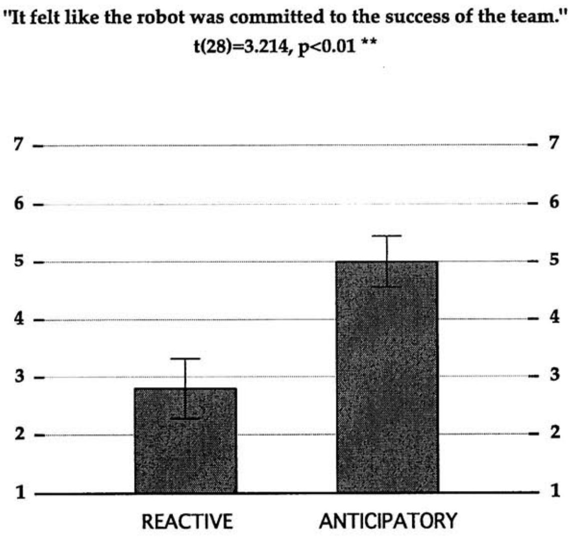
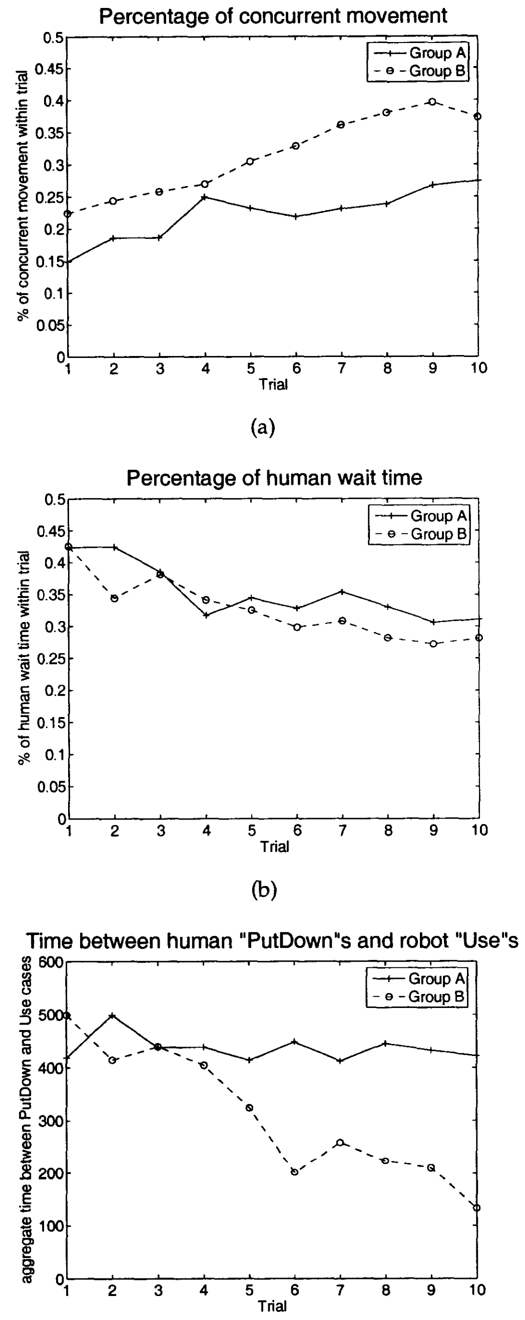


Figure 3-9: Significant self-report difference between participants in the *REACTIVE* (Group A) and the *ANTICIPATORY* (Group B) conditions, using a T-test for independent samples. Error bars indicate standard error.



(a)

(b)

(c)

Figure 3-10: Three measures of fluency per cart over ten trials, averaged over study condition A (REACTIVE) and B (ANTICIPATORY). (a) percentage of concurrent motion within trial; (b) percentage of human idle time; (c) aggregate time between human *PutDown* to robot *Use* delay.

teamwork are separate from those of simple task-time efficiency. If this is the case, we would like to discern possible quantitatively measurable causes for the above-mentioned perceptual differences.

However, while there is a large body of work measuring verbal fluency, there are no generally accepted measures of fluency in shared-location joint action, even for human teams. We therefore propose three fluency metric hypotheses, and compare the mean performance of the two groups along these measures in a post-hoc analysis of our study.

- *Hypothesis I: Concurrent motion* — In post-experiment interviews, some of our participants noted a sense that the team was well synchronized when "both team members were constantly in motion". We tested the hypothesis that the amount of human-robot concurrent motion was different between the anticipatory and the reactive condition. To do so, we measured the percentage of frames within each trial in which both human and robot were in motion (i.e. in transition between two location-based internal states), and indeed found those to be significantly different between the two groups (A: **0.227**; B: **0.322**; t(25)=3.11; $p < 0.005$). Figure 3-10(a) shows the mean percentage of concurrent motion for each of the 10 trials, averaged over subjects in each group. The graph shows that while the percentage of concurrent motion is improving for both groups, it does so at a higher rate in the anticipatory action condition.

- *Hypothesis II: Human idle time* — Our second hypothesis for a measure of fluency is the amount of time the human spent waiting for the robot. We postulated that if the human was to spend much time waiting, it would feel like the team was not working fluently. However, measuring the percentage of frames in which the human waiting (i.e. not doing anything, and not in transition between two location-based states), we found no significant difference between the two groups (Figure 3-10(b)). This was true for both the mean and the convergent human idle time. Both groups decreased the human waiting time at an approximately equal rate, and with similar results.

- *Hypothesis III: Functional delay* — We denote our third hypothesis "functional delay", i.e. we postulate that the amount of time passing between the human's action and the robot's consequent action was different between the two conditions. To test this, we measured the time between the human's *PutDown* action and the robot's subsequent *Use* action. We found this measure to be significantly lower for Group B (A: **436.78**; B: **310.64**; t(25)=5.04; $p < 0.001$), and more decidedly so for the second half of each subject's trial sequence, after the robot has adapted to the human's construction pattern (A: **432.07**; B: **205.08**; t(25)=6.28; $p < 0.001$) (Figure 3-10(c)). In the reactive case, there is virtually no change across trials.

While not ruling out additional factors, this evidence points in a promising direction with regards to possible quantitative measures affecting fluency in human-robot joint action. However, these findings are only initial and lay the groundwork for future research, in which each of these hypotheses needs to be separately controlled for, and evaluated for its effect on the human team member's perception of the robot's fluency, commitment, and task contribution.

3.4.3 Discussion

The open-ended segment of the post-experiment questionnaire reveals a qualitative difference between the two conditions. Several subjects in Group B noticed the anticipatory behavior and remarked on it positively, e.g.: "it was nice when [the robot] anticipated my next move", or "[the] robot's anticipation of my actions was impressive and exciting". Negative remarks in Group B usually referred to a desire for even more anticipatory behavior, such as "[the robot] could do better by getting the first tool before/while I take the first part, because it was a consistent process and could be predicted", or "the robot should watch what I'm grabbing in advance."

Somewhat surprisingly, many subjects in Group A — without having been informed that the study was related to anticipatory action or that the robot was meant to be adaptive — noted with frustration that the robot did not predict their actions. We view this tendency as indicative of the fact that adaptiveness and anticipatory action are natural expectations of a robotic teammate in a repetitive task. Quotes from Group A included: "I was hoping that the robot would learn to anticipate more", "I expected more predictive behavior from the robot", "[the] robot was not able to anticipate [the] human's actions", and "it might have been more efficient if after a few carts the robot could pick up on the order in which i was bringing in the parts and be prepared with the equipment to join it."

Group A's positive comments regarding the robot's performance were limited to remarks shaped by a low level of expectation from the agent: "The robot seemed to do what was expected", "the robot did not mess up", and "the robot was highly responsive and never let the human down with its predictability," were representative responses in this condition.

Notions of Teamwork

It is interesting to note that several subjects in Group A noted that the team felt "lopsided", that "the human was the one who strategized, the robot just sat there", that the human "was more important than the robot", and that "the team's performance was highly dependent on human innovation". Subjects in this group concluded that "the robot seemed more like an assembly tool than a team member", that they "didn't see the robot as a team player", that the robot was used "as a tool", and one subject said that they "didn't get a sense that the robot really cared about the success of the team." In contrast, in Group B only one subject noted that they "felt that the success or failure of the task was [their] responsibility." Conversely, one other stated that they "trusted [the robot] more over time, as it seemed to anticipate what [they were] going to do." The rest of the subjects in Group B did not address the balance of the team, the issue of trust, or that of commitment, in any way.

Effect of Repetition Size

As noted in Section 3.4.2, we believe that the relatively minor improvement in mean task efficiency through anticipatory action is related to the small number of repeating trials in the experiment. Appraisal of server logs, as well as user testimony, reveals that in many cases subjects experimented with various construction strategies in the first few runs, which caused the Bayesian model to converge more slowly. This seemed to be particularly true when subjects noticed that the robot changed its behavior, causing them to experiment with different construction sequences in an attempt to reveal the robot's modus operandi. One reason for this behavior was the experiment's insistence on identical instructions for both groups, not revealing that the robot would adapt to the human's consistent behavior. Several subjects explicitly noted that the team would have performed better had they known in advance that the robot learned to anticipate their actions. Another possible way to counter this effect would be to discount the learning over time (see also: Section 3.5).

Effect of "Best Score" Indicator

We also believe that the display of the game's all-time "Best Score" in the user interface was detrimental to the experiment as it might have caused subjects to experiment with different strategies instead of forming a consistent behavior pattern. Originally intended to motivate subjects to faster performance, the exceedingly good record time (only possible with a well-adapted agent) provoked subjects to question their strategy attaining a significantly worse score, and subsequently to change it several times over the course of the experiment.

3.5 Conclusion

In this chapter, we introduced a framework for evaluating shared human-robot fluency, and have presented a cost-based anticipatory action selection mechanism. We showed initial results on both the theoretical analysis of this method and its effect on untrained humans, showing significant differences in the subject's perception of the robot's fluency, commitment, and contribution, while showing only a small difference in mean and convergent task efficiency. In order to explain this discrepancy, as well as quantitatively evaluate the notion of fluency, we proposed three fluency metric hypotheses and compared these between conditions, finding significant differences along two of these metrics.

Several improvements to our method present themselves: in the discussed framework, the robot has no knowledge of the structure of the task. Domain-specific knowledge can decrease the action space at each decision point and fortify the accuracy of the probabilities of subsequent states.

We believe that our system can also be made more robust by introducing a discount factor in the learned state transition distribution, making more recent moves by the human teammate more salient to the robot.

Furthermore, the estimate of the human's turnaround time h should be state-specific and could be learned by the robot during the collaboration.

It also makes sense to evaluate the relative effect achieved by the state transition distribution learning, as opposed to the cost analysis during action selection. Also, the scalability of our method should be evaluated by increasing task complexity.

Additionally, the effects of anticipatory action vis-a-vis an expert — instead of a naïve — human teammate, is of interest, as is a controlled evaluation of the effects of the proposed fluency metrics on the efficiency of the task and the perceived fluency and commitment of the robot.

Chapter 4

A Perception-Action Cognitive Architecture for HRI Fluency

In the previous chapter, we investigated the probabilistic concept of anticipatory action. Wishing to represent this concept as part of a perceptual symbol system, we now describe our core cognitive architecture — inspired by the neuropsychological principles and findings outlined in Section 2.1 — modeling concepts as perceptional symbols and allowing for a dynamic processing stream through the establishment of top-down perceptual simulation and emulation. We also describe the relation between perceptual symbols and actions, and show how action selection is biased via the change in perceptual parameters.

Grounded in perceptual symbol theories of cognition, and in particular that of simulators [Barsalou, 1999], we model concepts as perceptual processes and biases, residing within the perceptual system rather than in an amodal centralized semantic network.

4.1 Modality Streams

The highest-level organizing principle of a perceptual modality is a *Modality Stream* in which *process nodes* filter perceptual signals from the world (through the sensory layer) towards action (Figure 4-1). These process nodes can be thought of as akin to Damasio's *convergence zones* [Damasio, 1989, Simmons and Barsalou, 2003].

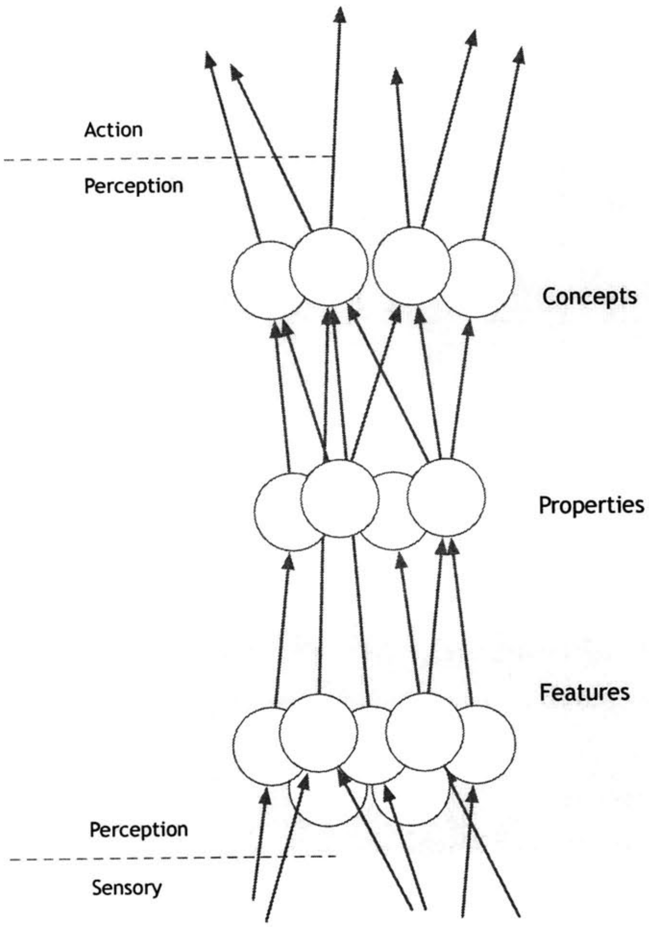


Figure 4-1: Schematic of a modality stream

Within each modality stream, there exist process nodes of three types, in order of distance from the sensory layer: *features*, *properties*, and *concepts*. These types of nodes are elaborated upon later in this chapter. Connections between nodes in a stream are not binary, but weighted according to the relative influence they exert on each other. Similar weights also affect the top-down processing influence between various nodes.

In addition, inspired by neurological findings, process nodes in a certain modality's perception stream are not isolated, and can be connected to nodes in different modality streams and to action nodes. It seems reasonable to believe that in most cases these inter-modal and perception-action connections occur on the concept level. However, we have found it useful in some cases to directly connect feature and property nodes across modalities.

4.2 Process Nodes

The basic data structure in a modality stream is the *process node*, shown in Figure 4-2. A process node may correspond to a certain perceptual process (such as finding speed of a tracked object), may "react" through activation to a certain perceptual feature (such as the existence of a certain color), or both.

Process nodes have a set of outgoing (*downstream* or afferent) connections and a set of incoming (*upstream* or efferent) connections. The direction of the arrows in the diagrams in this thesis corresponds to the bottom-up perceptual processing from sensory activation to concept and action. However, it is important to note that efferent activation also flows in the network in the opposite direction to the diagram arrows.

Each process node contains a floating-point activation value, α, which represents its excitatory state, may affect its internal processing, and is in turn forwarded (potentially altered by the node's processing) to the node's downstream connections. This downstream activation is multiplied by the weight of the connection through which the activation travels.

A separate simulated activation value σ is also taken into account in the node's activation behavior and processing (see: Section 4.3). This happens in two ways: a dampened simulation value $\sigma \times f_s$ — where f_s denotes a simulation-forwarding factor — is added to the activation propagation when a node activates its downstream processing nodes. In addition, $\sigma + \alpha$ is used as a motor action trigger value in the so-called *action nodes* described in more detail below.

The basic downstream operation of a process node is shown in Code Segment 1. Whenever an incoming connection activates the node, it processes the incoming activation values *a*

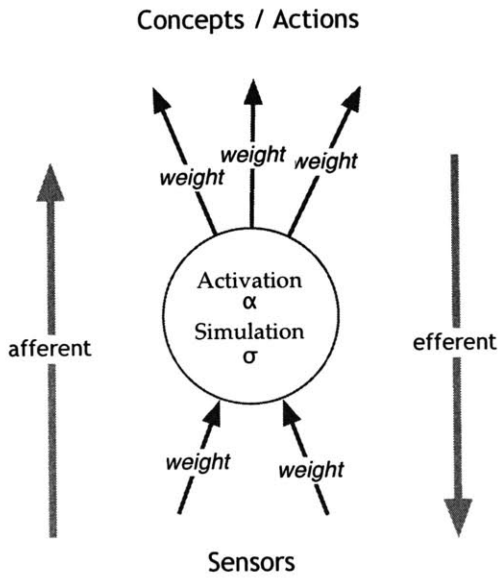


Figure 4-2: Schematic of a process node.

and data d coming in from its incoming connections and adds the processed activation a' to its own activation value α. This value, in combination with the processed data d' is then used to activate to its downstream connections. The particular decision on altering the node's processed activation value a' and possibly altering the incoming data is determined in each node's INTERNAL-PROCESS method, reflecting the node's behavior.

Whenever a node is not activated, its internal activation value α decays linearly, and the residual activation value α is used to activate downstream connections without additional data.

4.3 Simulation

Downstream activation alone does not enable us to model the simulated perception caused by efferent connections, which is considered to exist in the human perceptual system. To enable top-down processing and to model the bidirectional nature of the cognitive archi-

Code Segment 1 Afferent activation of a process node.

```
ProcessNode-Activate($node, a, d$)
if $a = 0$
  then return
$[a', d'] =$ Internal-Process($a, d$)
$\alpha \leftarrow \alpha + a'$
$down\_act = \alpha + \sigma \times f_s$
for each $node'$ in $outgoing\_connections$
do ProcessNode-Activate($node', out\_connection\_weight(node') \times down\_act, d'$)
```

tecture attempted in this work, we must implement a mechanism to simulate perception from the concept down towards the sensory system.

According to this principle, the activation of a concept can evoke the construction of past and fictional perceptual snapshots. Citing Barsalou [1999]: "Productivity in perceptual symbol systems is approximately the symbol formation process run in reverse. During symbol formation, large amounts of information are filtered out of perceptual representations to form a schematic representation of a selected aspect. During productivity, small amounts of the information filtered out are added back."

The mechanism enabling the top-down processing of perceptual memory is the efferent connection between process nodes within a certain modality. In this framework, both activation and data are assembled from higher-level processing nodes and collected in lower-level perceptual processes, influencing their activation and data handling.

4.3.1 Design considerations

Enabling top-down processing through simulation raises two key design questions:

The first is whether a separate simulation mechanism is necessary at all, or whether simulation could be modeled within the normal perceptual activation framework.

We have found the single-stream approach insufficient for two reasons: if nodes are connected both upstream and downstream with symmetrical connections, simulated perception would result in an excessively high activation of perceptual processing nodes. We wanted to allow for simulated activation to be appropriately dampened in comparison to sensor-derived activation, when appropriate.

Furthermore, when activation works both in the afferent and the efferent direction at the same time, the system risks falling into a feedback loop in which downstream activations would immediately activate the upstream nodes that feed them, deranging into constant full-fledged activation unrelated to incoming sensory data.

We have therefore opted for a separate activation and simulation process, which uses two kinds of connections enabling the system to flow dampened simulated activations in parallel to full sensor-derived activations. This might also be in line with current findings in neural activation, which identifies separate afferent and efferent activation streams in the human brain [Barsalou, 2007].

The second question was whether a node needs to have a single activation value, and therefore a simulated activation affects the same activation value as a sensor-derived activation. This would mean that simulated activation is internally indistinguishable from a sensor-derived activation, and would imply that — within our system —simulating an experience really *is* equivalent to experiencing it.

While empirical evidence in neuro-psychology is inconclusive as to that question, we believed that it was worthwhile to explore this "strict" variant of simulated perception. However, while we were able to model a simple task (as the one in Chapter 5) using this approach, we found it insufficient for a more complex modeling task (as the one in Chapter 6). For that reason, we have added the simulation value σ to the modeling of the process node as described above.

In sum, Code Segment 6 shows the efferent activation (simulation) function of a process node.

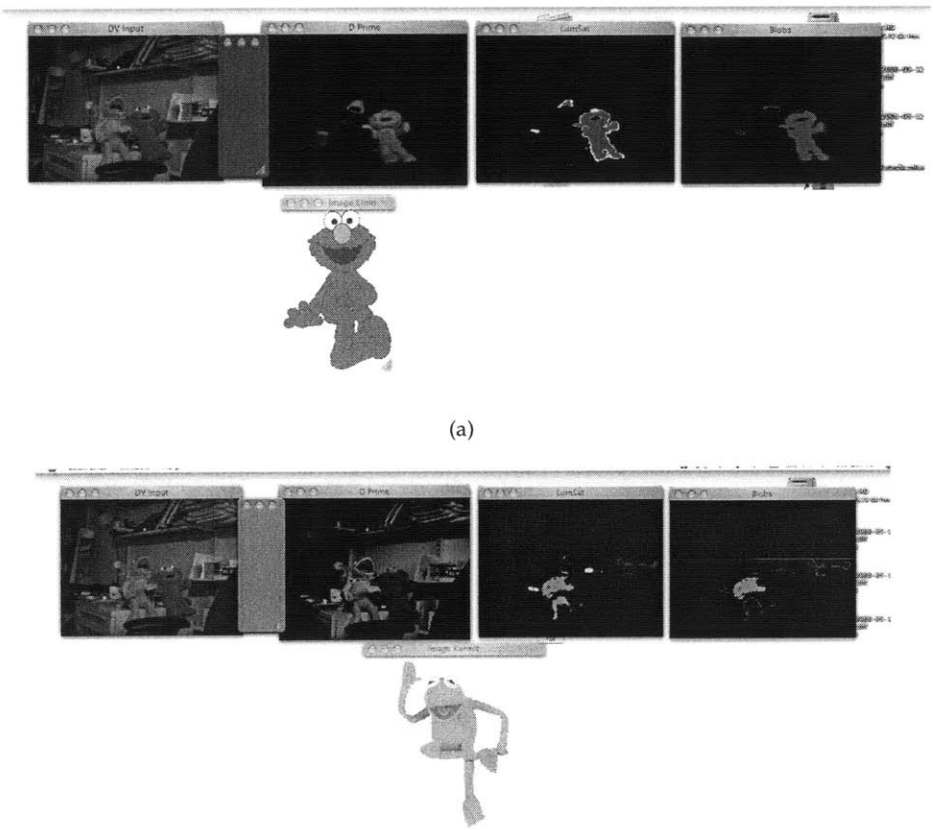
Code Segment 2 Efferent simulation of a process node.

```
ProcessNode-Simulate($node, s, d$)
if $s = 0$
  then return
$[s', d'] =$ Internal-Process-Simulate($s, d$)
$\sigma \leftarrow \sigma + s'$
for each $node'$ in $incoming\_connections$
do ProcessNode-Simulate($node', in\_connection\_weight(node') \times s', d'$)
```

4.3.2 Top-down processing

A central claim of this work is that fluency in human-robot interaction can be achieved by simulated perception biasing the the dynamics of the perceptual system through top-down processing. Figure 4-3 exemplifies this approach in a simple inter-modal simulation example: an auditory percept (in (a), the sound "Elmo") activates a canonical visual memory of the figure, which — using the same pathway utilized in visual perception — detects the dominant color of the image. This color is then used as a bias affecting the low-level visual buffer, shifting it towards detection of similarly colored areas, eventually priming the system to detect the Elmo puppet in the visual field more easily. For comparison, (b) shows the same visual processing map when an auditory "Kermit" is used to prime the perceptual system.

In the architecture described herein, simulation affects perceptual activation by lowering the bottom-up perceptual activation necessary for the activation of various process nodes. This in turn lowers the actual sensory-based activation threshold for action triggering, resulting in increasingly automatic motor behavior based on simulated data. This principle will become clearer when elucidated by the implementation examples brought forth in the following two chapters.



(a)

(b)

Figure 4-3: Top-down processing and cross-modal activation in a perceptual-based associative memory model.

4.4 Types of Process Nodes

We classify the process nodes in our systems into four categories: Sensory, Feature, Property, and Concept nodes. The distinction between three perceptual categories (*Feature*, *Property*, and *Concept*) should be thought of as a conceptual categorization rather than a functional one, as there is no inherent difference between the behavior of nodes in each of these three categories. Sensor Nodes, on the other hand, are special in that they initiate the activation propagation triggered by the world events to which they serve as an interface.

Sensor Nodes are activated by external world events, and are the only nodes that are explicitly enumerated by the system for activation.

Feature Nodes represent the first-level feature extraction from the sensory activation. They usually generate a data packet associated with their activation.

Property Nodes are second-level detection nodes, which operate on feature data. Their activation represents a single property, the extent of which is indicated by the level of their activation.

Concept Nodes are the highest level intramodal process nodes. Their activation is based on a pattern of property node activations, and they usually represent a specific modality's "gateway" to action nodes, as well as to other modalities (akin to the *Modality CZs* in Simmons and Barsalou [2003]).

4.5 Action Nodes and the Action Network

Similar to the modality stream model, actions are also organized along activation nodes, called *action nodes*. These reside in a so-called *action network*.

While in most ways similar to the process nodes described above, action nodes are subject to additional management by the action network since they access a joint and exclusive resource — the motor system. In the past, several models have been offered for motor control arbitration, in robotics [Brooks, 1991] as well as in animated characters [Perlin and Goldberg, 1996, Burke et al., 2001].

This work arbitrates motor control in the following manner: each action node is associated with a motor behavior, as well as a set of joints affected by its behavior. Additionally, action nodes are grouped into *exclusivity cliques*, denoting a mutually exclusive set of actions that cannot overlap.

Each action has an activation threshold above which it will trigger the motor behavior associated with it. Importantly, in order to determine activation triggering, both the current activation value α and the current simulation value σ are taken into account, as described earlier in this chapter. Actions can be "one-way" or "return", meaning that an action behavior can play out from the current joint configuration to a new one, or it can be one that automatically returns to its original joint configuration.

Some actions are "binary" in nature, which means that, once they have been triggered, they will move from their start to their end position, and possibly back, regardless of the activation value. Other actions are "variable", moving to an extent of their behavior that's proportional to their activation value. In any case, if the action node activation drops below a second threshold, the action will either cease (for "one-way" actions) or return to its original state (for "return" actions).

When two or more actions, which are not in the same exclusivity clique, attempt to control a joint, the joint's behavior is blended by the action network into an average of these joint requests, much as described in Burke et al. [2001].

When an action is activated while another action in its exclusivity clique is running then, if the new action's activation is higher than the one currently running, it will abort that action's behavior. If its activation is lower, the new action request is delayed until its activation supersedes the currently running action.

A central claim of this thesis is that perceptual processing and action are not separate systems, but part of the same inter-effective mechanism. This principle is implemented in the fact that the activation process does not distinguish between perception and action nodes, and the same connection-weight logic is applied to both kinds of nodes. Moreover, simulation also involves both perception and action, enabling a motor action to bias perceptual processing if it is activated, be it to a full extent, or only partially as part of an inactive motor simulation.

4.6 The Update Loop

Code Segment 3 shows the top-level update loop of our perceptual symbol system. Each sensory input activates the appropriate sensor node, which in turn propagates this activation according to its afferent connections. Similarly, simulated perceptions activate simulation chains in the efferent direction. Finally, in the action network update, each action node gets an opportunity to adjust the joint position based on its current activation value, and those of other action nodes.

Code Segment 3 Main update loop.

```
MAIN-UPDATE(sensor_input, simulation_input)
for each si in sensor_input
do sensor ← FIND-SENSOR(modality_streams, si)
    data, value ← GET-DATA-VALUE(si)
    PROCESSNODE-ACTIVATE(sensor, value, data)
for each si in simulation_input
do node, data, value ← GET-NODE-DATA-VALUE(si)
    PROCESSNODE-SIMULATE(node, value, data)
ACTIONNETWORK-UPDATE()
```

Figure 4-4 shows a visualization of an active perception-action system as used in the implemented cognitive network described in Chapter 6.

In the architecture laid out above, practice operates as follows: as expectations are increasingly formed (through mechanisms similar to those described in Chapter 3), simulated perceptions prime the process nodes in the perceptual streams and enable less perceptual input to result in earlier action activation. This corresponds to the proposition that practice moves action activation from deliberate and slow to automatic and fast processes. Moreover, we achieve this result through a perceptual simulator approach, as well as through the integration of perception and action.

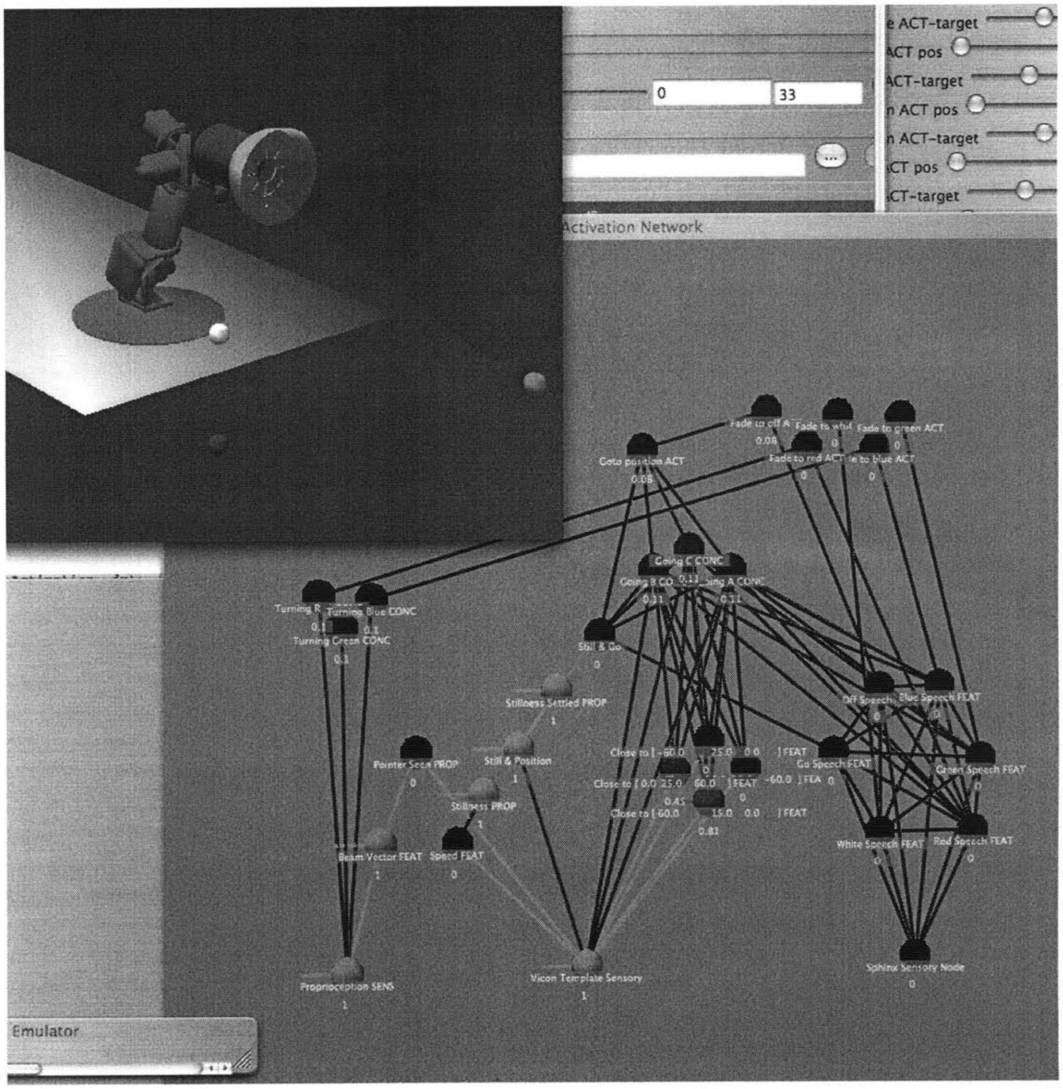


Figure 4-4: Screenshot from the software running the perception-action network described in this chapter, as used in the implementation in Chapter 6

4.7 Connection Reinforcement

An additional mechanism of practice in our system is that of *weight reinforcement* on existing activation connections. While most node connections are fixed and set by the designer of the perception-action network, some can be assigned to a connection reinforcement system, which will dynamically change the connection weights between the nodes.

This refinement of connection weight can be conceived to be influenced by a number of factors, such as frequency of co-ocurrance, the contingency between the two nodes, attention while creating the perceptual memory, and affective states during perception[1]. Roy and Pentland's work in multimodal concept learning suggests insight into the potential acquisition of these intermodal connections [Roy and Pentland, 2002].

In the implementations described in the following chapters, connection reinforcement works according to the contingency principle, with the aim of reinforcing connections that co-occur frequently and consistently, and decreasing the weight of connections that are infrequent or inconsistent.

More formally, let a_s denote the activation value of the upstream (or *start*) node of a connection governed by the connection reinforcement mechanism, and a_e the activation value of the downstream (or *end*) node of the same connection. Then, if a_s is greater than a threshold α, we compute $\delta_{s \to e} = (a_s - \alpha) \times (a_e - \beta)$, for a second threshold β. This value $\delta_{s \to e}$ is then added to the current connection weight.

Let $\delta_s \equiv max_e(\delta_{s \to e})$ and $E_s \equiv argmax_e(\delta_{s \to e})$. E_s is thus the downstream node of node s that applies for the highest reinforcement value at a given moment. In each update, we additionally decrease the connection weight of all connections $s \to e$, for $e \neq E_s$, by δ_s.

The result of this approach will reinforce consistent coincidental activations, but inhibit competing reinforcements stemming from the same source node. The effects of this approach are empirically analyzed in Section 6.4.1.

[1] The role of attention and affective state on the retention of learned or practiced mechanisms could explain why certain marginal lessons are retained for a long time if they occurred in a highly attuned or traumatic setting

4.8 Design Principles

Our cognitive design can be viewed as a hybrid approach between a connectionist model, based on activation values, and an information-based or algorithmic model. While each process node models a certain excitatory value, it is not equivalent to a strict neural representation of a single scalar. Instead a node represents a higher-level logical operation through its reaction to, and processing of, incoming data and the forwarding of a manipulated copy of said data to its downstream connections. So, while — on a high level — we model our system according to neural perceptual streams and inspired by neuropsychological models of brain region segmentation, we do not emphasize low-level neurological fidelity, but instead use authored algorithmic processes within each of the process nodes.

This stems from a belief that a biologically inspired, yet computationally authored approach is appropriate for robotic systems, as they are situated and embodied, and thus need to operate in real time and with noisy sensory data. This is particularly true when the explicit aim of this work is achieving a sense of fluency and speed in the robot's behavior. In addition, our model makes authorship — which we argue is implicit even in so-called "pure" connectionist models — explicit, and allows the HRI system builder to take full advantage of algorithmic design, instead of painstakingly designing around this ostensible limitation.

While predominantly intended to empirically evaluate the principles of perceptual simulation in repetition practice for HRI as laid out in this and the following chapters, our architecture can also serve as a basis for future development of similar human-robot teamwork systems. The following principles may guide the designer of such a system:

Action network: Enumerating and authoring the actions the robot is capable of is usually the first step in the cognitive network design. The robot's potential actions should then be classified into "binary" and "variable" actions, as well as into "one-way" and "return" actions.

Modality streams: Usually one modality stream should be planned for each sensor of the robot. A proprioceptive modality stream is often easy to implement, and should be defined as well if task-appropriate. This modality stream keeps track of current joint positions of the robot.

Afferent connections: Downstream connections define relations within a modality stream, and default to $weight = 1$. In cases where two or more connections enter a processing node, their weight should by default be divided by the in-degree of the node. In such a node, the designer can bias the importance of a certain incoming connection by distributing the 1-sum weights in an unbalanced manner.

Efferent connections: Upstream connection weights usually default to a dampened value of the corresponding downstream connections. Both the initial site of simulation and the depth of simulation needs to be taken into account by the network designer. Particular attention is to be paid to the reconstruction of lower-level perceptual features (data) by the activation of higher level elements.

Inter-modal reinforcement: In cases of inter-modal reinforcement, the initial connection weight between nodes in different modalities is usually small (later increased by the reinforcement system). The main design task in this item lies in the selection of nodes which are candidates for this kind of connectivity, and the direction of influence between those nodes.

Simulator/Emulator: Simulation can be triggered based on a number of events, such as long-term memory correlations, or anticipatory processes based on learning. The choice of learner for the simulation/emulation mechanism is independent of the proposed architecture, and in the work described in the following chapters, we used a simple Bayesian learner on a single-sequence Markov chain. However, alternative mechanisms are equally appropriate.

Calibration: Real-world data should finally be used to calibrate the connection weights, as well as the simulation coefficients used in the top-down processing mechanism.

Chapter 5

Perception-Action Architecture: Pilot Experiment

The first field test of the cognitive architecture described in the previous chapter was in the form of a "Patty Cake" game between a humanoid robot and a human, in which the robot practices the timing of its actions to meet the human's hands, which move in a repetitive sequence. For readers not familiar with English nursery rhymes, Patty Cake is a clapping game popular in many countries, which "alternates between a normal individual clap with two-handed claps with the other person. The hands may be crossed as well. This allows for a possibly complex sequence of clapping that must be coordinated between the two." [Various Authors, 2007].

While the Patty Cake trials were initial and only involved a limited spectrum of perception and action, we believed that they provided a good testing ground for core architectural issues related to the overall research program. Indeed many central issues and theoretical questions arose during this stage, which will be discussed towards the end of this chapter.

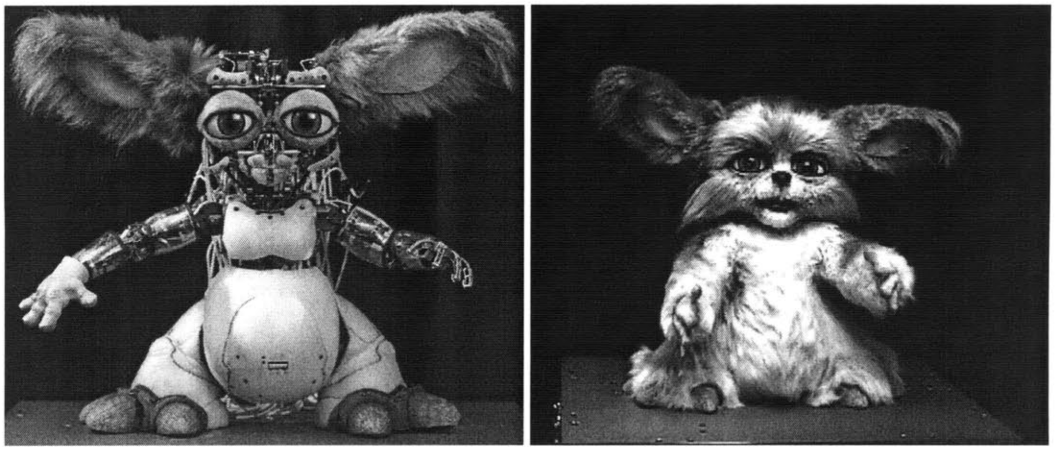


Figure 5-1: "Leonardo", the expressive humanoid robot used in this study. The left picture shows the robotic structure, the right picture shows the robot when cosmetically finished.

5.1 Platform

We applied the cognitive architecture described in the previous chapter to an expressive humanoid robot, Leonardo ("Leo"), shown in Figure 5-1. Leonardo is a 65-degree of freedom (DoF) fully embodied humanoid robot that stands approximately 2.5 feet tall. It is designed in collaboration with Stan Winston Studio to be able to express and gesture to people as well as to physically manipulate objects. The robot is equipped with two 6-DoF arms, two 3-DoF hands, an expressive (24-DoF) face capable of near human-level expression, two actively steerable 3-DoF ears, a 4-DoF neck, with the remainder of the DoFs in the shoulders, waist, and hips.

The robot's sensory input device was a Vicon motion capture system which was used to identify and track the location of the human's left and right hands. The motion tracking system implemented uses ten cameras with rapidly-strobing red light emitting diodes to track small retroreflective spheres that can be attached to clothing and other materials. The positions of these trackable markers can be determined with very high accuracy within the volume defined by the cameras - often to within a few millimeters of error at a tracking rate of 100-120Hz. In this experiment, markers were fitted onto gloves in two distinct patterns. These markers were then pre-processed to provide the system with a 3D location of each of the human's hands.

Since the robot used for this experiment is not designed for physical contact, we have implemented a slightly modified version of the game, in which the robot's and the human's hands meet at a slight horizontal distance. This does not affect the timing, perception, or motor trajectory aspects of the interaction. However, it could be argued that the tactile and auditory perception of the clap can and should be used as an additional signal with which to coordinate the joint behavior.

5.2 Perception-Action Network

The perception-action network that was built for this experiment is displayed in Figure 5-2. Two feature nodes parse the raw input from the Vicon sensor and detect the 3D position of the left and right hand, respectively. These feed into a property node each for the detection of the left and right hand *up*ness property. A third property node is activated according to the overall upward speed of the motion in the sensory field.

The property nodes in turn feed into three concept nodes: one representing the concept of a left hand Patty Cake clap move, one for the right hand move, and one for the move involving both hands clapping.

Note the insertion of the conjunction (∩) operator between the left and right hand *up*ness properties and the two-handed game move. In the afferent mode, this node activates only when all of its incoming connections activate. While this abstraction could have been made in the node's INTERNAL-PROCESS method, we believe this to be a general enough mechanism to afford a separate process node type. This naturally raises the question of whether relationships between perceptual nodes (or Convergence Zones) are implemented in the network's structure alone, or whether they have separate representations.

The concept nodes are connected to the two action nodes, triggering the appropriate action in the game.

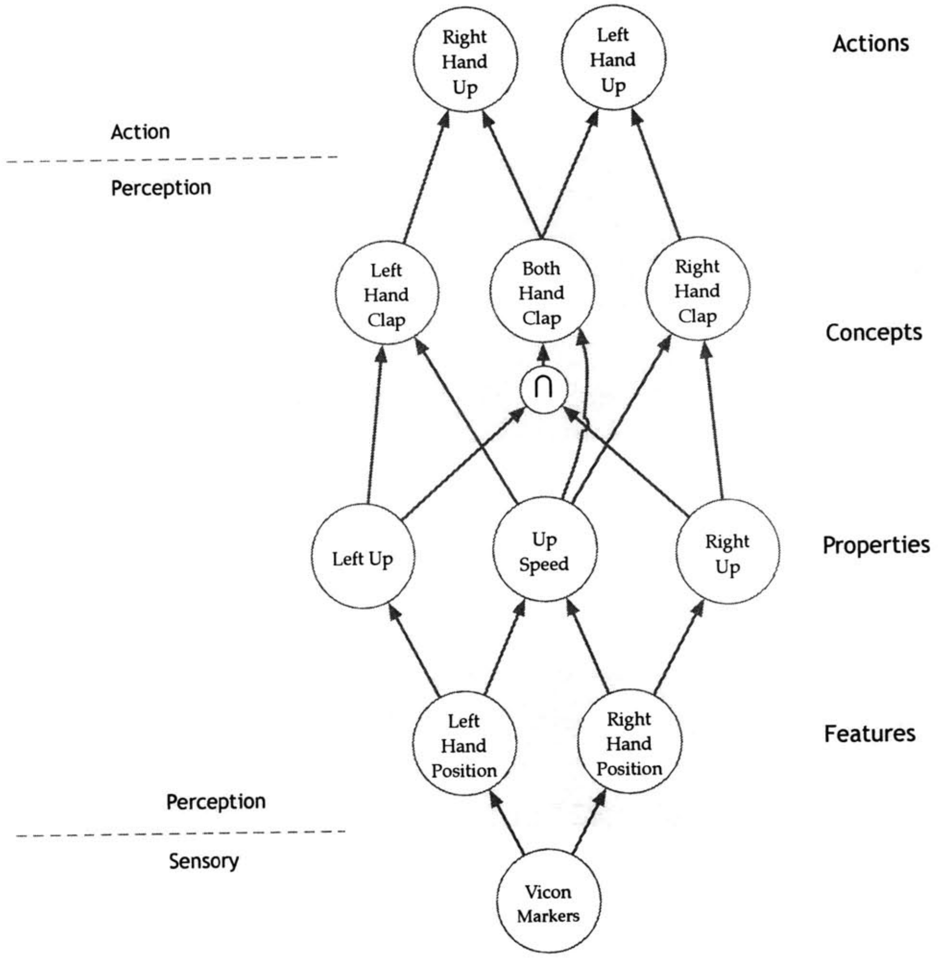


Figure 5-2: Perception / Action Nodes for the Patty Cake game

5.3 Internal Processing

The internal processing for the feature nodes converts the "visual field" to 3D locations of each of the hands, and activates if the appropriate marker pattern was successfully detected. It generates a data packet with the location, which is fed through the downstream activation into the property nodes.

Code Segment 4 Internal processing of a **hand position** feature node.

```
INTERNAL-PROCESS($a, d$)
  $d' \leftarrow$ FIND-HAND-POSITION($d$)
  return [$a, d'$]
```

The up speed property node activates according to the average one-step derivative of vertical movement of the left and right hands. Negative movements are capped at zero.

Code Segment 5 Internal processing of the **up speed** property node.

```
INTERNAL-PROCESS($a, d$)
  $next\_d \leftarrow$ VERITCAL-PROJECT($d$)
  $speed \leftarrow next\_d - last\_d$
  $last\_d \leftarrow next\_d$
  return [$speed$, NIL]
```

The property and concept nodes are value-only, and do not pass on any data besides their activation. The activation value of the left and right "up" property nodes is proportional to the vertical location of the hand in the visual field, whereas the concept nodes do not perform any internal processing.

Code Segment 6 Internal processing of an ***upness*** property node.

```
INTERNAL-PROCESS($a, d$)
  $upness \leftarrow$ VERITCAL-PROJECT($d$)
  return [$upness$, NIL]
```

An interesting result of this network layout, given that activation is additive, is that the faster the up-motion that is detected (by the activation of the "up speed" property node), the lower the hand needs to be to activate the appropriate concept node. Some of the anticipatory timing necessary to meet the human's gesture at the right moment thus emerges without being explicitly calculated.

5.4 Anticipation and Rhythm

Perceptual simulation in this pilot experiment is triggered at the concept node level. We then follow a notion of perceptual simulation that does not evolve into a full sensory experience, but instead operates only down to the property level, which can be thought of as parallel to holistic Convergence Zones [Simmons and Barsalou, 2003]. Thus, based on which of the concept nodes is activated, either the left, the right, or both *up*ness property nodes are active.

Simulation of Patty Cake move concepts are triggered by a combined *anticipation* and *rhythm* module (Figure 5-3). This module is an extension of the anticipatory action predictor discussed in Chapter 3. However, there are several differences between the initial version of the predictor and this one.

To recap, in the earlier version the agent acted on the next perception anticipated by the Baysian learner whenever the previous step was complete. There was no aspect of timing inherent to the event. In this extension, the learner also tracks the beat, or rhythm, of the interaction triggering the appropriate perceptual simulation at the time at which the next event is expected.

Second, while in the original anticipatory action algorithm there was a simple one-to-one mapping between anticipated perception and subsequent atomic action, in this version the anticipated concept is instead used to generate a perceptual simulation which integrates with sensor-based perception to possibly generate action. This follows the notion of top-down perception more closely, as simulated perception is truly integrated with sensor-based perception.

Finally, we have increased the length of the Markov chain history on which the Bayesian predictor is learned from one step to three steps. The memory size of the Markov chain can be generalized to an arbitrary number.

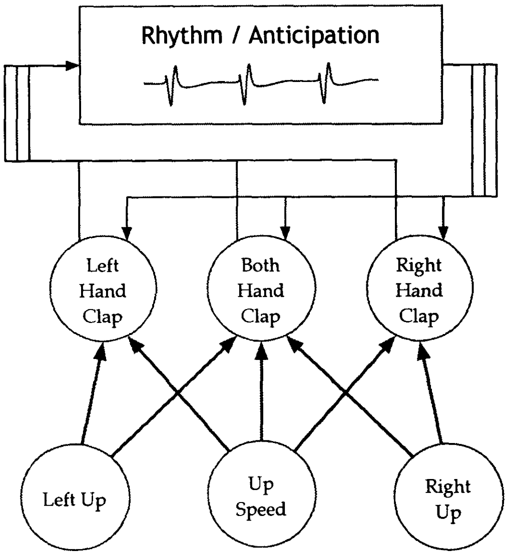


Figure 5-3: Simulation is triggered by a rhythm module.

In sum, during the game of Patty Cake, the agent captures the periodicity of the human's moves by refining a distribution over a three-step Markov chain, as sensed in the appropriate concept nodes. At each step, if the expected move at time t is denoted by σ_t, and $\{t^1, \ldots, t^n\}$ are the times of the n previous recorded moves, the agent generates a probability distribution over the set of possible next moves

$$p_{i|j^1 \ldots j^n} \equiv P(\sigma_t = i | \sigma_{t^1} = j^1, \ldots \sigma_{t^n} = j^n)$$

where $i, j^1, \ldots j^n \in \{Left, Right, Both\}$.

Similarly to Section 3.3, we use a naïve Bayesian estimator with a Laplace correction of 1 [Kohavi et al., 1997]. A change to move $\sigma_t = i$ following moves $\sigma_{t^1} = j^1, \ldots \sigma_{t^n} = j^n$

increases the counter $n_{i|j^1\ldots j^n}$. $p_{i|j^1\ldots j^n}$ is then estimated by

$$p_{i|j^1\ldots j^n} = \frac{n_{i|j^1\ldots j^n} + 1}{\sum_{k=1}^{|Moves|} n_{k|i|j^1\ldots j^n} + 1}$$

5.4.1 Rhythm

In parallel, the rhythm of the game is estimated using a simple linear filter with a constant decay, measuring the time between the previous and the current move. If the time that has passed since the last move is within a certain distance ϵ (expressed as a fraction of the beat) from the expected beat length, the appropriate concept node is triggered in upstream simulation.

5.4.2 Combined Rhythm / Anticipator

Code Segment 7 shows the process of rhythm detection and anticipation, or simulation triggering. In this listing, b is the current estimated beat length, α is the filter coefficient, and if there is no move input from the concept node layer, σ_t is NIL.

Code Segment 7 Anticipating the next human move.

```
RHYTHM-ANTICIPATOR($t, \sigma_t$)
if $|(t - t^1) - b| < b \cdot \epsilon$
   then $[i, p] \leftarrow$ MARKOV-ESTIMATE($\sigma_{t^n}, \ldots \sigma_{t^n}$)
        PROCESSNODE-SIMULATE($node_i, p$)
if $\sigma_t \neq$ NIL
   then $b \leftarrow b \cdot \alpha + (1 - \alpha) \cdot (t - t^1)$
        $\sigma_{t^n} \leftarrow \sigma_{t^{n-1}}$
        $\vdots$
        $\sigma_{t^1} \leftarrow \sigma_t$
```

As can be seen in the above code segment, the simulation activation is proportional to the probability of the expected move. In the pilot implementation, MARKOV-ESTIMATE returns only the probability of the top expected move, although a possible alternative would be to activate all the concept nodes with their respective probability.

5.5 Simulation

The result of a simulated perception triggered by the rhythm/anticipation module is a limited activation of the concept and property layer nodes that correspond to the perception of the anticipated next human move. The rhythm component of the anticipator triggers this activation at a time that matches the currently perceived rhythm of the interaction.

Since simulated activation acts as an additive component in the afferent perceptual stream, the outcome of this simulation are twofold: first, lower hand positions and slower speeds are sufficient to trigger the robot's own actions, making its behavior increasingly anticipatory the more confident the Bayesian model of the human's actions are. In a constant-speed behavior of them human, the robot will react to the human's motion at an earlier point in the trajectory. This results in a move from a purely reactive behavior (the robot starting its action only when the human has reached the vertical threshold position) to a meshing behavior (the robot moving at the same time as the human and increasingly meeting the human's move at its endpoint).

Second, since the robot's simulated perception results in partial activation of concept nodes, and as a result of action nodes - the robot begins the motion trajectory associated with the appropriate action ahead of time, and thus puts its body into a more favorable position to complete the gesture when the full perceptual event is experienced through sensory-based activation.

5.6 Experiment

Figure 5-4 shows the effects of a repetitive practice game of Patty Cake with Leonardo. Frames (a) and (b) demonstrate the game range, in this case for a human LEFT / robot RIGHT move. In (a) both the robot and the human are in their respective pre-move idle positions. (b) has both players at the end of their move trajectory. The lines on the left side of the frames indicate the robot's thumb location in the idle, full-up, and a "move onset" position, which was defined as a pose distinctly separable from the idle position.

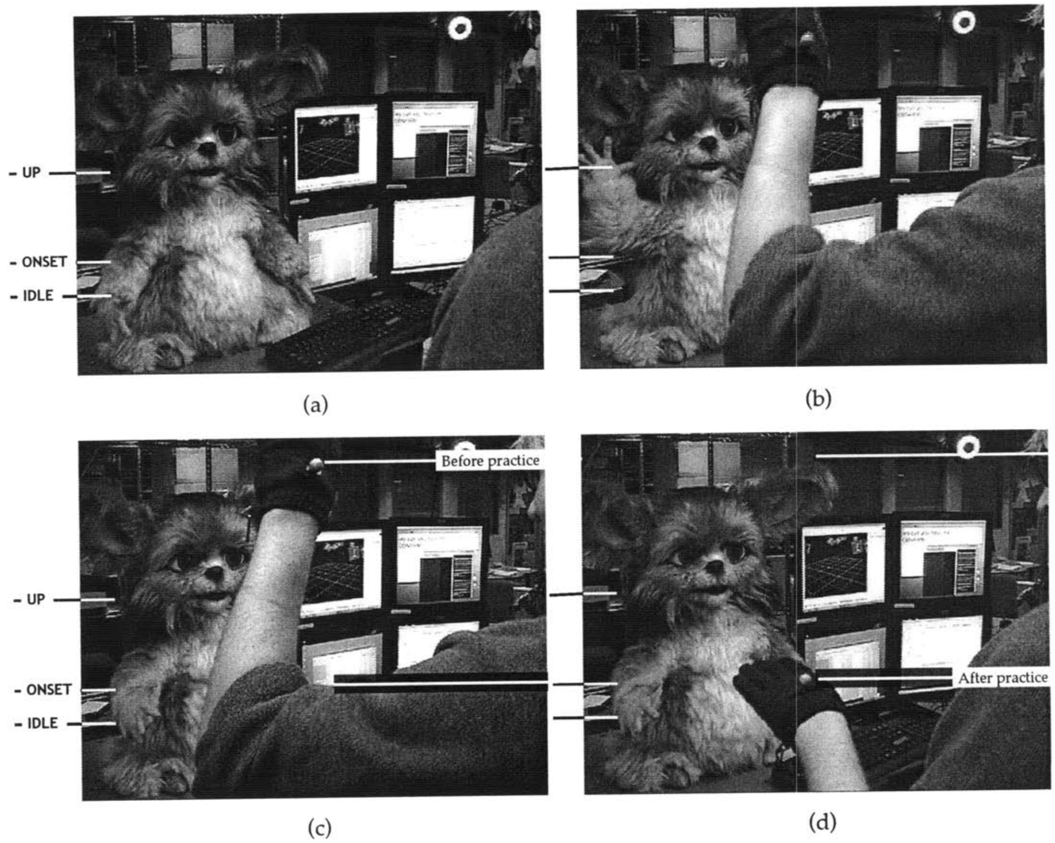


Figure 5-4: Practice effect in human-robot Patty Cake game (human LEFT move matching robot's RIGHT move); two-step repetition over 7 iterations. (a) shows the robot in idle position, (b) shows both players at the end of their trajectory. (c) shows the moment of the robot's move onset at the beginning of playing the game. By the time the robot starts moving its hand up, the human has already completed his full trajectory. (d) shows the robot's move onset after a few rounds of repetitive practice: the robots starts moving soon after the onset of the human's move, making it more likely for the robot to meet the human's move at the end of their trajectory.

In this experiment, the human chose a two-step game sequence alternating the left and right hand move repeatedly.

Frame (c) shows the first time the left hand move is employed by the human (step 2). The robot's move sets on only after the human's move is completed, demonstrating a fully reactive behavior. Frame (d) shows the robot's behavior six iterations later (step 14). The robot's move onset, triggered by the combination of simulated perception induced by the rhythm/anticipator and the sensory-based perceptual activation, happens quite early in the human's move.

This results in a more meshed behavior, enabling the robot to meet the human's move almost simultaneously. Additionally, this behavior leads to more concurrent movement, complying with the empirical findings in Section 3.4.2.

5.7 Discussion

While the Patty Cake game represents an indubitably simple manifestation of the proposed architecture, including only three separate actions and a single perceptual modality, it still proved to be an interesting pilot implementation of the approach put forth herein.

The implemented architecture was shown to model a number of important mechanisms inherent in fluent joint action. The relationship between anticipation and simulation was demonstrated as we increase the generality of the anticipatory framework of Section 3.3. We showed how simulated perception and sensory-based perception can combine to result in early action, longer concurrent motion, and shorter delay between the human's and the robot's action — metrics which we have earlier suggested to play a significant role in humans' sense of fluency.

We also showed how speed and position can be combined in a unified activation framework, resulting in an emergent behavior of appropriate timing.

Some research questions arise from this first implementation, which set the path for further investigation:

5.7.1 Simulated activation vs. Sensory-based activation

One of the first considerations in implementing this system was whether simulated activation is actually equivalent to real activation. Moreover, if it results in the same activation from another source, how does it integrate with sensory-based activation? Is it a parallel system that 'copies over' values from simulation to perception, is it additory, does it compete with sensory activation? As discussed above, in this case, we have implemented a system in which simulated activation is equivalent to real activation, additively changing the same variable of each node. This approach was altered in later implementations of the system, in which simulation was separate from — but in interaction with — sensor-based activation, as described in the previous chapter.

Through this implementation, we have also realized that there is a risk of uncontrolled feedback loops if perceived and simulated activation operate in completely identical ways, and have opted to separate these into two separate pathways. Each node has upstream and downstream outgoing connections (which — in this case — are fully symmetrical, but an asymmetrical implementation was used for our second experiment), and while perception activates only the downstream connections, simulation only operates on upstream connections. This has been found to stabilize the network.

This seems to be supported by some preliminary evidence that the right hemisphere processes bottom up information and the left hemisphere generates it top-down [Barsalou, 2007].

5.7.2 Actions and Simulation

Another issue that arises with the proposed approach is how actions activate as part of the simulatory pathway. Since simulation is initiated at the concept level, how do we prevent actions to trigger regardless of sensory-based perceptions? In the discussed implementation it is a matter of degree of activation, using an additory paradigm between perceived and simulated activation on the property level. However, we later found it useful revise our approach, considering a system that "knows" which activations are simulated and which are not.

5.7.3 Processing Symmetry

If a node does some processing in addition to its function as a pathway node (for example it might threshold a certain incoming activation), this processing must be different between the efferent and the afferent direction. It would be a worthwhile research path to determine whether there is a generalizable way to inverse processing between sensory-based and simulation-triggered perceptual activation.

5.7.4 Competing Simulations

How should the system deal with competing simulations? In this Chapter's implementation we have proceeded in a winner-takes-all approach, although a parallel simulation could also be considered. In this case, is there a need for an inhibitory mechanism for competing simulations?

5.7.5 Length of Simulation Streams

A further important question is how far upstream a simulation activates. In this pilot study, we found that the correct behavior arises if simulation moves up to the property node level. A sliding target approach might be necessary, moving simulations further into the sensory processing stream if needed.

5.7.6 Extensions

The proposed experiment could be further extended by generalizing the Up property nodes to action extent detection properties, which can be used to determine how much of a partner's action has been executed. This, in combination with the speed node might be enough to model Time-to-Impact necessary for meshing joint actions. The evaluation of completion-percentage of a partner's action could conceivably be learned for novel actions.

In addition, the original version of the Patty Cake game includes individual claps between each of the joint claps. These highly embodied signals might well have a crucial role in coordinating timing between the players.

The inclusion of an auditory and tactile signal resulting from the clapping gesture (both the individual and the joint clap) could also prove useful in both the rhythm estimation and the triggering of actions at the right time.

Chapter 6

A Collaborative Lighting Task

Building on the first generation implementation described in the previous chapter, and in an attempt to design a network that includes multiple modalities and a more complex network, we devised a collaborative task in which a human and a robot can work together to achieve a joint goal, and in which practice plays a role in improving human-robot team efficiency and fluency.

6.1 Robotic Platform

The robot employed in this evaluation was AUR, a robotic desk lamp. The lamp has a 5-degree-of-freedom arm, a variable aperture that can change the light beam's width, and a ColorKinetics LED lamp which can illuminate in a range of the red-green-blue color space. Chapter 9 describes of the design process and mechanical makeup of AUR.

AUR is stationary and mounted on top of a steel and wood workbench locating its base at approximately 36 inches above the floor. Its processing is done on a 2x Dual 2.66GHz Intel processor machine located underneath the workbench. Figure 6-1 shows the robot and the workbench.

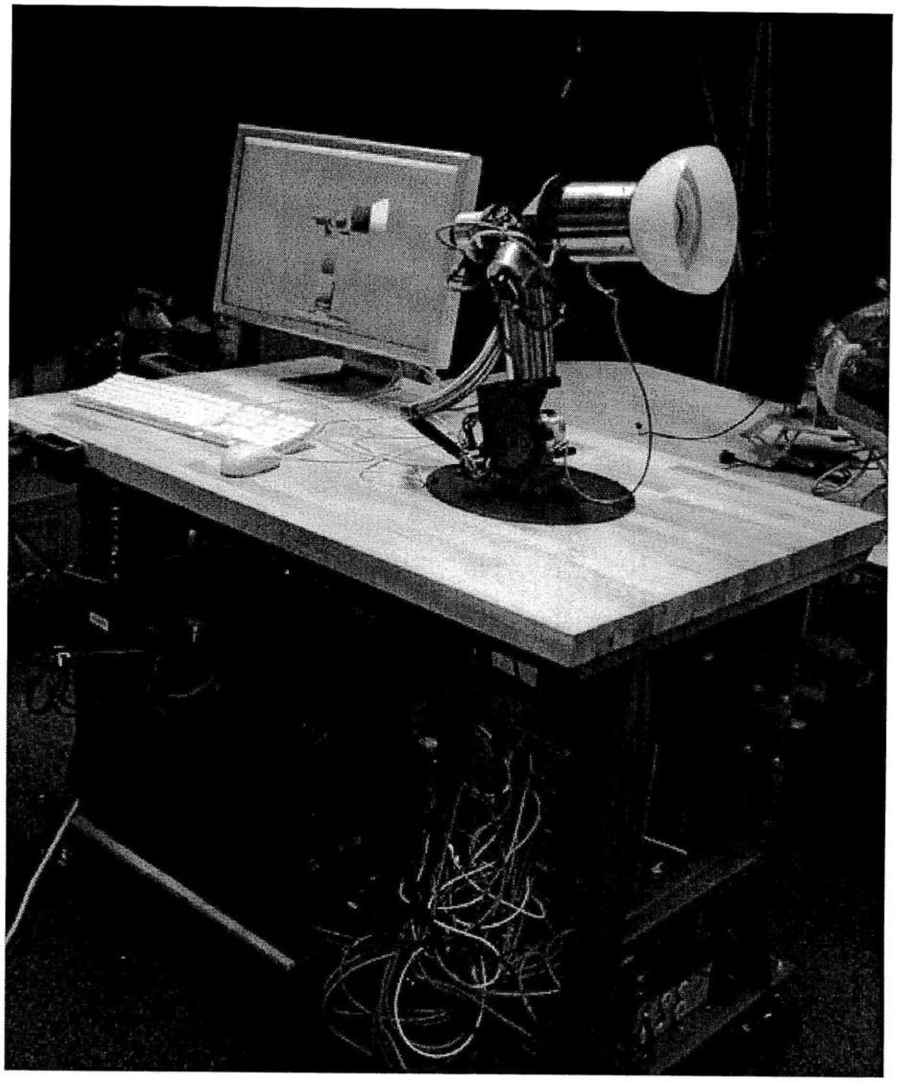


Figure 6-1: The robot AUR on its workbench

6.2 Task Description

The task was designed so that in order to achieve the goal, both human and robot needed to do the appropriate actions, and that the sequence of actions necessary would be repeated in a way that practice could improve performance on both the human's and the robot's part.

Additionally, we have attempted to design a task that is appropriate for a robotic desk lamp, but is strictly reproducible in a controlled experiment. The task was planned with the intent that it is relatively difficult for the human to learn the sequence of actions they need to do, and that there are parts of the task that only the human can do and parts that only the robot can perform.

In this human-robot collaboration, the human operated in a workspace as depicted in the diagram in Figure 6-2 and the image in Figure 6-3.

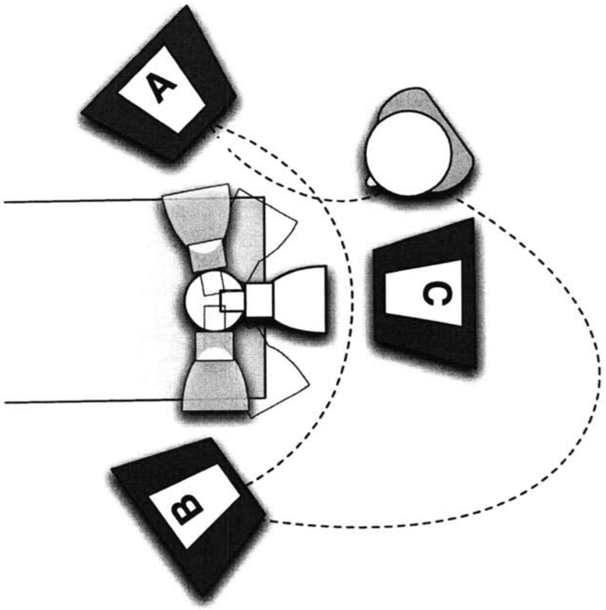


Figure 6-2: Diagram of the collaborative lighting task workspace

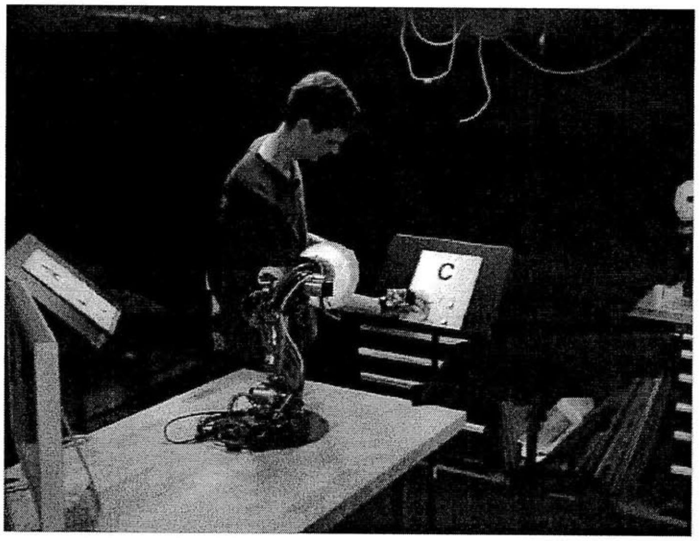


Figure 6-3: Photograph of the collaborative lighting task workspace

The robot, from its tabletop vantage point, could direct its head to different locations around its own axis, and change the color of the light beam. When asked to "Go", "Come", or "Come here" the robot would move to the location of the person's hand, assuming the hand was relatively static. Additionally, the color changed in response to speech commands to one of three colors: "Blue", "Red", and "Green"'.

To complete the task, the person was asked to stand in front of the lamp. The workspace contained three locations (A,B,C). At each location there was a black music stand with a white cardboard square labeled with the location letter, and including four doors (Figure 6-4). Each door, when lifted, revealed the name of a color written underneath (Red, Blue, or Green).

The task was to complete a sequence of 8 actions, which was described in diagrammatical form on a sequence sheet as shown in Figure 6-5. This sequence was to be repeated 10 times, as quickly as possible.

Each action in the sequence specifies: a general location A, B, or C, and an indication of which of the four doors to open. The action is completed when the lamp shines the specified color of light at that location. This would result in the sound of a buzzer, indicating the

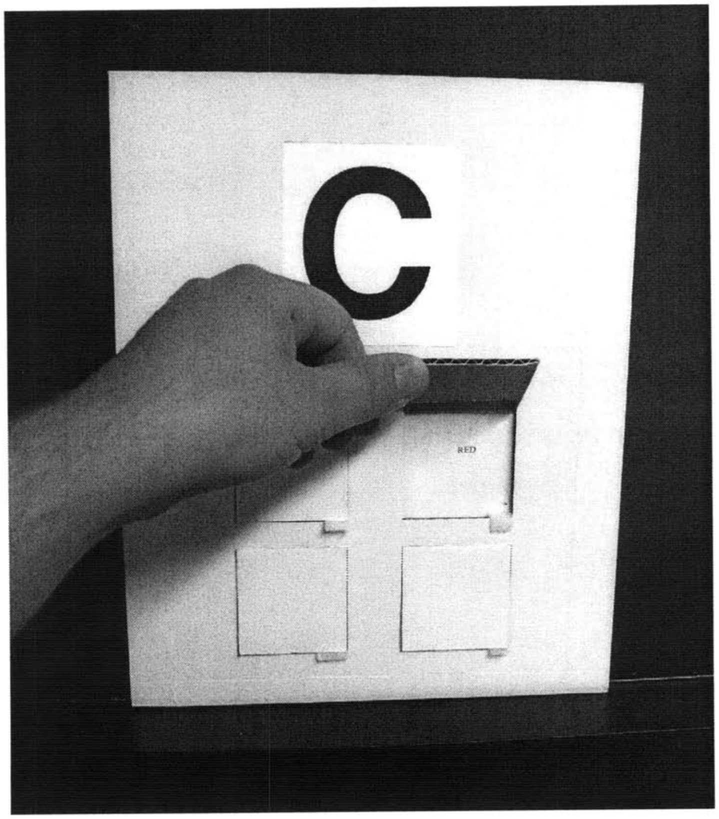


Figure 6-4: Cardboard used in the lighting task. Each location is indicated by a letter A–C and has four doors, each of which hides the name of a color. Different sequence patterns have different colors under each door location.

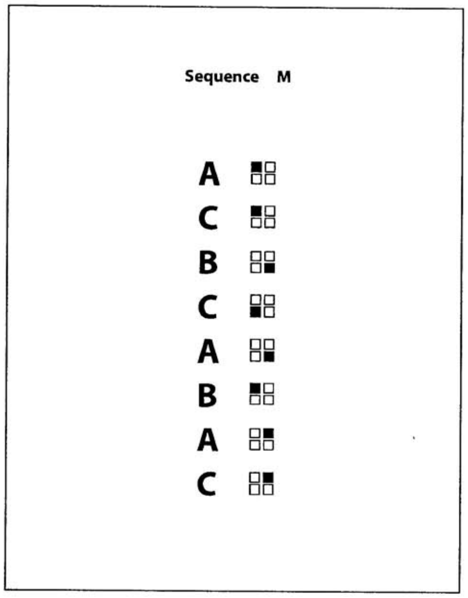


Figure 6-5: Sample experimental sequence

person should move to the next action in the sequence. A different buzzer was sounded when a whole sequence was completed.

6.2.1 Sensing

The robot's sensory input device was a Vicon motion capture system which was used to identify and track the location and orientation of the human's right hand and head. The motion tracking system implemented uses ten cameras with rapidly-strobing red light emitting diodes to track small retroreflective spheres that can be attached to clothing and other materials. The positions of these trackable markers can be determined with very high accuracy within the volume defined by the cameras - often to within a few millimeters of error at a tracking rate of 100-120Hz. In this experiment, markers were fitted onto a glove and a headband in two distinct patterns. These markers were then processed to provide the system with a 3D location of the human's hands and head. We recorded the location of the hand, as well as the location and orientation of the head to disk at a frequency of 10 times per second.

The system also takes input from Sphinx-4, an open-source, Java-based speech recognition system created by the Sphinx group at Carnegie Mellon University, in collaboration with Sun Microsystems Laboratories, Mitsubishi Electric Research Labs, and Hewlett Packard [Walker et al., 2004]. The commands recognized by the system in this task were: "Come", "Come Here", "Go", "Red", "Blue", "Green", and "Off".

6.3 Cognitive Network

To solve the task described in this chapter, we designed the cognitive network depicted in Figure 6-6, based on the elements described in Chapter 4.

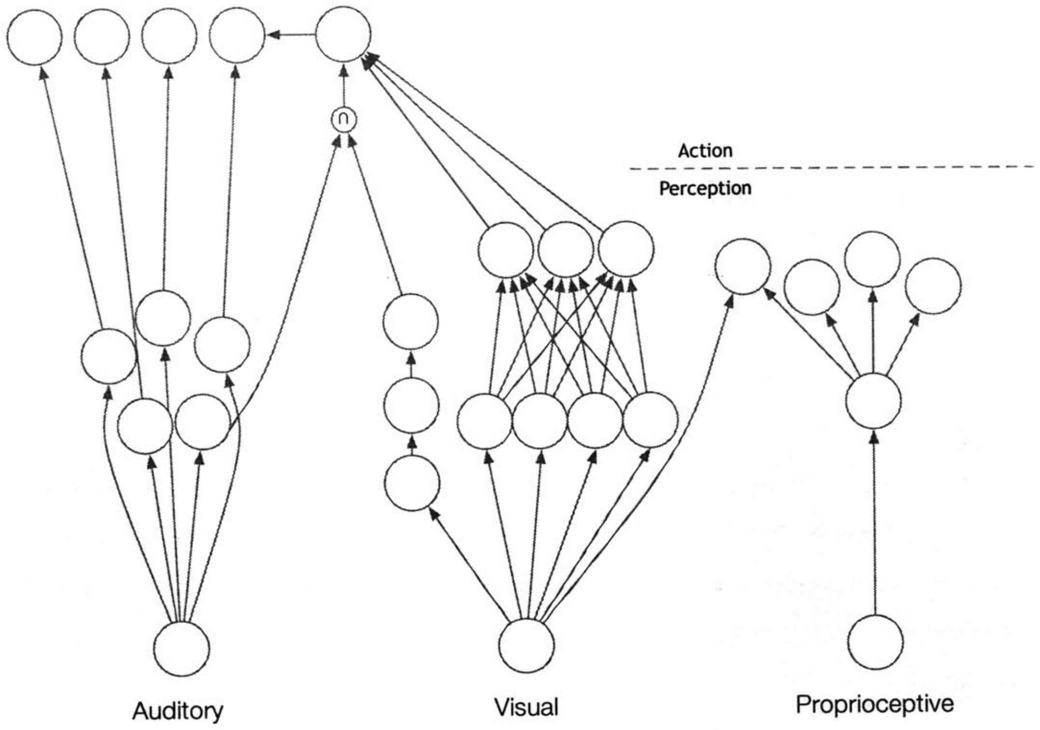


Figure 6-6: The cognitive network used in the light following task.

The network is made up of three modality streams, a visual, an auditory, and a proprioceptive. The action network includes five actions, three color changing nodes, a color "off" node, and a "goto position" node.

6.3.1 Visual Modality

The visual modality stream is depicted in Figure 6-7. Its sensory input stems from the Vicon motion capture system, indicating the hand position in the workspace.

Two feature areas are downstream from that Vicon sensory node: in one, four workspace segmentation feature nodes activate proportionally to the proximity of the hand to each of the four corners of the workspace. In the other, a speed node detects the speed as a single-frame position derivative of the hand position. Downstream from the speed node, an inhibitory connection feeds the "Stillness" feature node, which simply detects the inverse of the speed node clamped between 0 and 1.

The stillness node feeds into an aggregator property node detecting the settling of the stillness feature. This node is a slow following (low-pass filtered) node with an instant negative response. It thus indicates whether the hand has been still for some consecutive period of time.

Three concept nodes represent the location of the hand near one of the task targets, A, B, or C.

The target concept nodes are connected to the "goto position" action node, as is a conjunction node combining the "stillness settled" and the "Go" speech feature node. A combination of a hand position near the target, an aggregate stillness of the hand, and a "Go" command will trigger this action, leading the robot to move its beam towards the appropriate position.

6.3.2 Simulation in the Visual Modality

Simulation in the visual modality is most relevant to the connection between the concept nodes and the workspace segmentation feature nodes. Over time, as the robot correlates

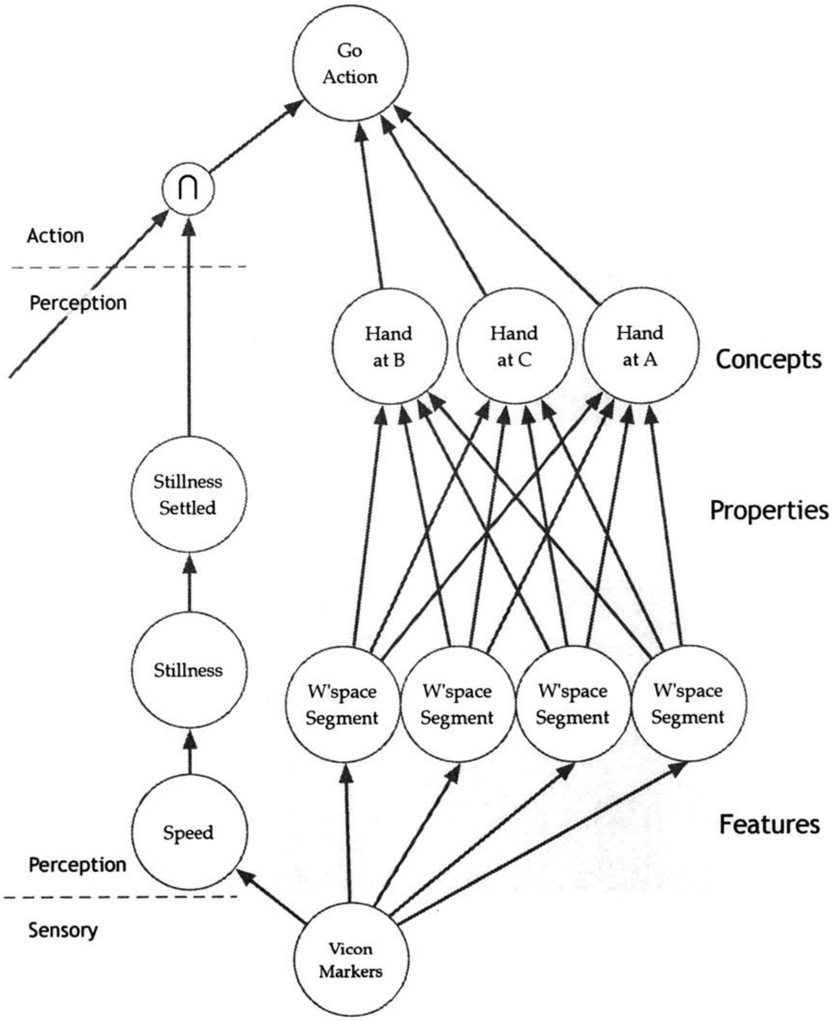


Figure 6-7: The visual modality stream used in the light following task.

the location of the hand near a target, both the afferent and the efferent connections between these two layers get reinforced.

Then, in anticipation of the next game step as determined by the simulator/emulator described below, the appropriate concept node, as well as the "Go" action node get a certain simulated activation, as determined by the action and concept simulation coefficients and the anticipatory probability. At this point, the reinforced efferent connections to the feature nodes cause a bias towards the perception of the hand in the appropriate workspace regions.

During such simulation, the workspace segmentation feature nodes are partially activated, and as a result can reach full activation with sensory data that is further away from each of the target positions. This results in a pre-step bias of this perceptual stream leading to anticipatory action on the robot's part.

6.3.3 Auditory Modality

The auditory modality stream is depicted in Figure 6-8. Its sensory input stems from the Sphinx speech recognition system, parsing the auditory input into speech tokens.

This sensory node feeds five speech feature nodes, which respond to the detection of a particular speech token in the auditory stream. As mentioned in the previous section, the "Go" speech feature feeds into a conjunctive node with the "Stillness settled" visual property triggering the "Go" action only when the hand position is settled. This prevents jitter on the part of the robot induced by slight variations in the hand position.

The four other speech feature nodes have afferent connections to the four light change action nodes. Note also that the "Go" action node causes the light to turn off to prevent light to shine while the lamp is in motion.

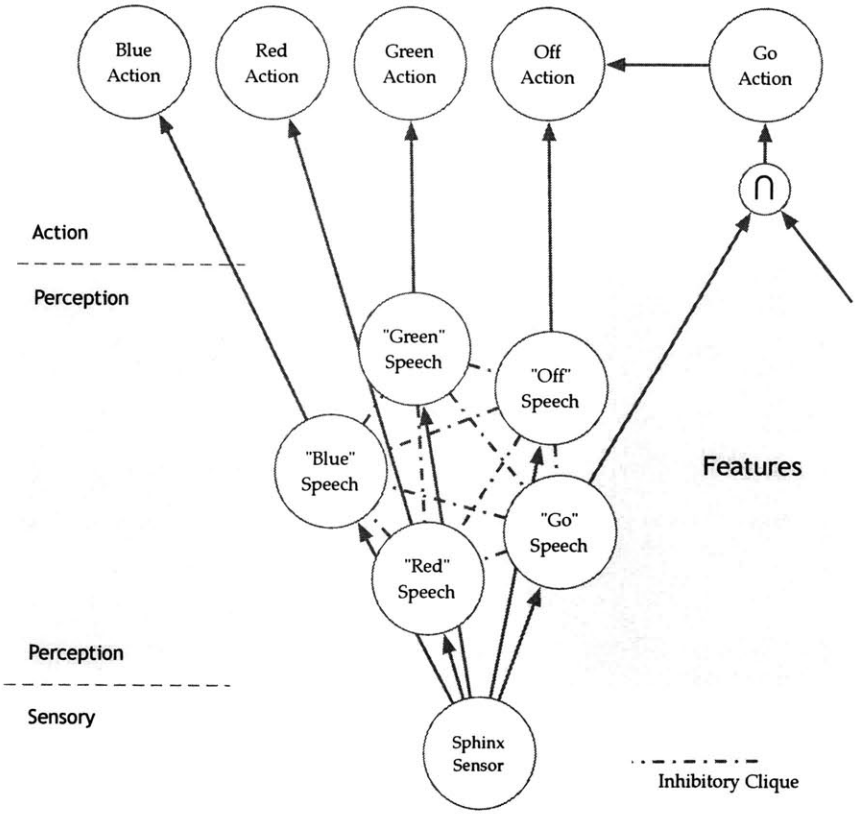


Figure 6-8: The auditory modality stream used in the light following task.

Inhibitory Clique

All speech detection feature nodes reside in a so-called *inhibitory clique,* which governs their relative exclusivity. This is not the same as the *exclusivity clique* described as part of the action network. The inhibitory clique does not disable other nodes in the clique, but each pair in this clique has a strong inhibitory connection, resulting in a new speech detection to minimize the activation of competing speech feature detections.

6.3.4 Proprioceptive Modality

Finally, the proprioceptive modality stream is depicted in Figure 6-9. Its sensory input is based on the joint positions of the robot, thus representing the robot's sense of physical self.

This simple modality stream has a direction feature node, which calculates the lamp head direction from the joint positions. This node, in turn, feeds into left, right, and center property nodes, which classify the overall orientation of the robot. In addition, and in combination with the hand position sensory node, a "towards hand" property is calculated from this feature (but not used in the experiment described below).

Inter-modal connectivity

The orientation property nodes in this modality are connected to the speech feature nodes in the auditory modality, and managed with the connection reinforcement mechanism described in Section 4.7. The aim of these connections are to enable the robot to learn a perceptual correspondence between its orientation and a speech detection perceptual stimulus. These connections are initially extremely weak, but can later be embellished by the connection reinforcement mechanism, as described in the previous chapter.

Thus, while there is no direct connection between the proprioceptive modality and the action network, this modality indirectly affects the color actions through their connectivity to the auditory modality.

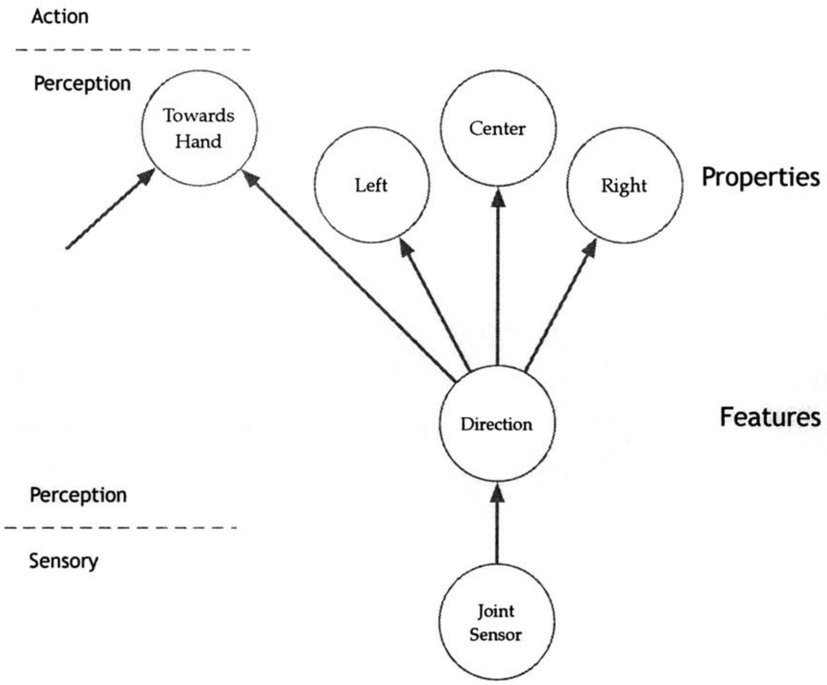


Figure 6-9: The proprioceptive modality stream used in the light following task.

6.4 Emulator

To model practice in this network, we designed a Bayesian sequence learner along the same lines as described in Section 5.4. Our Markov model used a 3-step sequence history.

At each step, after a few rounds of practice, the sequence learner estimates the probability of the appropriate hand-target concept, and "go" action being expected. This triggers simulation of both the appropriate concept and the action node.

For the concept simulation, we have heuristically found that a quadratic derived value results in a more appropriate learning curve over time. For the action simulation, we found a linear simulation value adequate.

The result of this emulation, in combination with the connection reinforcement described above, is that the workspace segmentation feature nodes simulate to the extent that they are correlated with the appropriate hand-target concept. Thus an increasing distance between the hand position and the correct target is adequate to trigger the appropriate response.

6.4.1 Analysis

Figure 6-10 shows the probabilities of anticipated concepts for a consistent human teammate, and their respective simulation values. The graph shows these probabilities at the same step for each attempt for a given expected concept, and for a not-anticipated concept.

Figure 6-11 shows the associated simulation value with the probabilities shown in Figure 6-10. As stated above, the concept was simulated with a quadratic derived value $s_c = (p + \gamma) \times p$, in this case with $\gamma = 0.4$. Compare this with a linear derived activation reaching the same settled value. For the action simulation factor, we used a linear coefficient, $s_a = p \times \gamma$. In this case, $\gamma = 0.25$

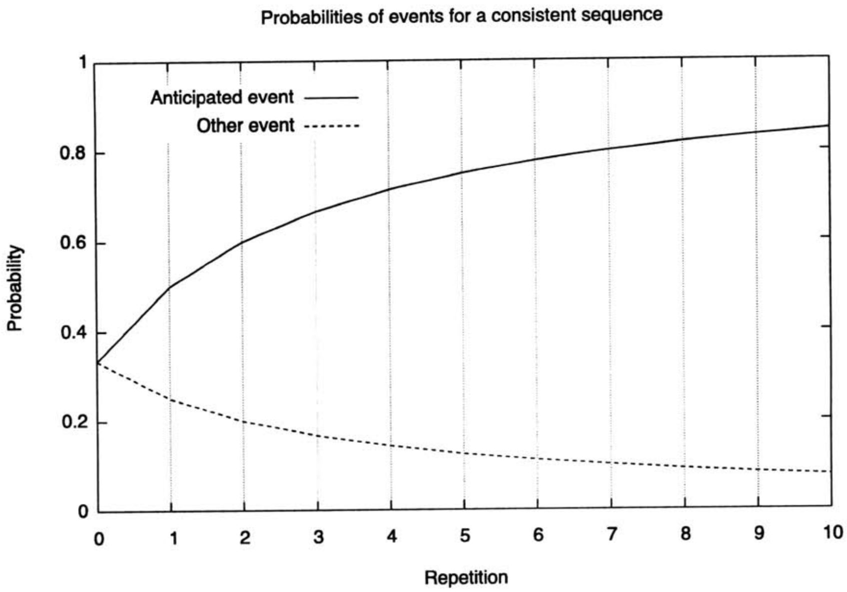


Figure 6-10: Probability for an anticipated event and a not anticipated event in case of a consistent human.

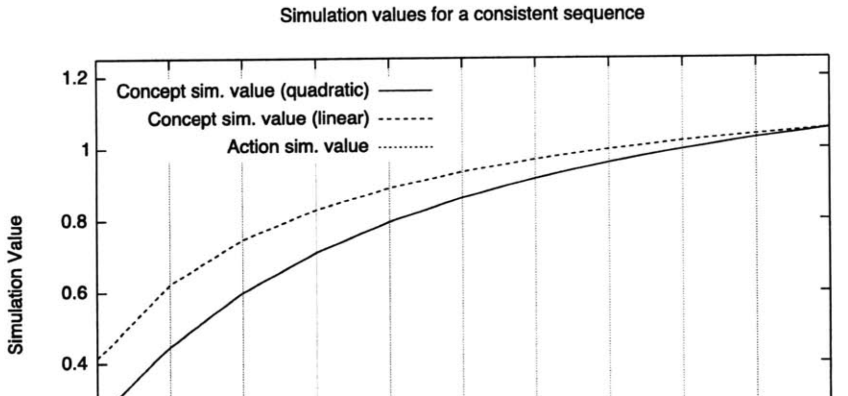


Figure 6-11: Simulation values for anticipated concept and action with a consistent human teammate.

Figure 6-12 shows the the change in trigger area for each of the areas using a simple A-C-B sequence with ten iterations. The longer the practice session, the more 'primed' the perceptual stream is, and the further away the hand position can be to result in a concept and subsequent action activation. This effect can also be seen graphically in Figure 6-13 for the same recorded data as above.

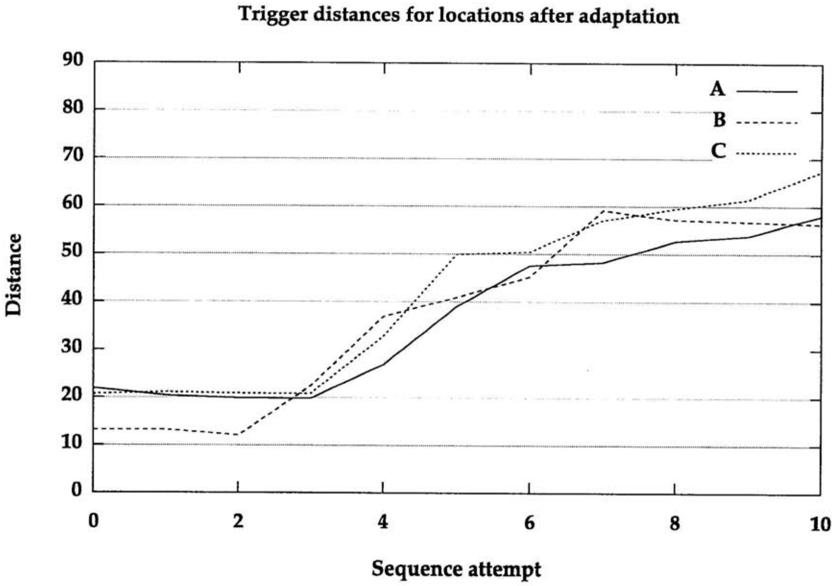


Figure 6-12: Effect of emulator on trigger distances for each of the three locations using a simple A-C-B sequence.

One interesting result was that if the human moves in the wrong direction for the next step, in many cases the robot is triggered to move briefly in the correct direction before following the human's command. This often results in a joint matched movement to one and then to the other direction performed by both the human and the robot. We believe that such an embodied mirroring behavior could play a role adding to the team's sense of bond, as well as to the human's perception of the robot as similar to themselves.

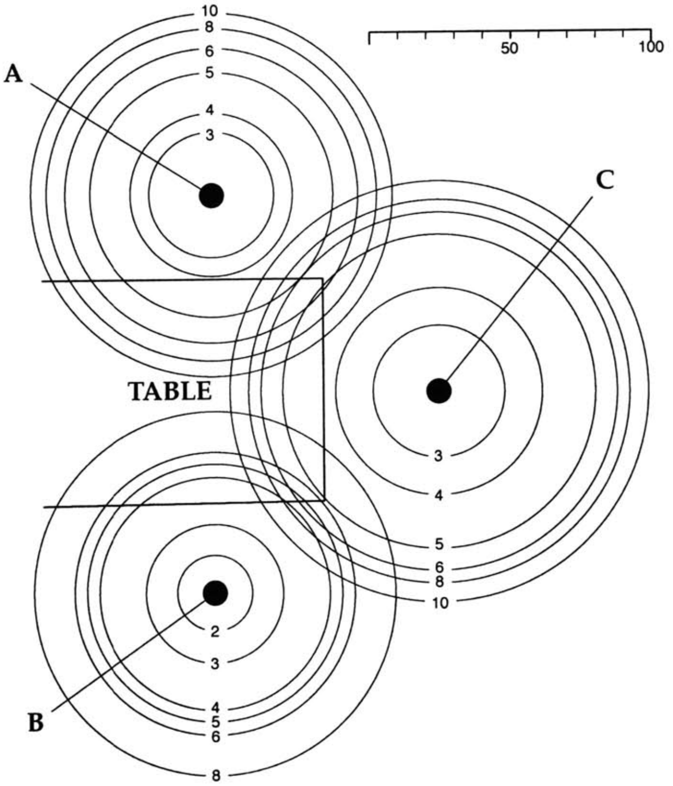


Figure 6-13: Effect of emulator on trigger distances for each of the three locations using a simple A-C-B sequence. Numbers indicate sequence attempts.

6.5 Inter-modal Simulation

In addition to the visual emulation, we used inter-modal simulation between the robot's proprioceptive property nodes ("Left", "Right", and "Center") and the speech feature nodes (see: Section 6.3.4). This corresponds to a repeating stimulation of a certain word in the auditory stream when the robot has a certain position, resulting in the perceptual simulation of that speech segment every time the robot returns to that position.

If there is a consistent correlation between position and color, the robot will increasingly trigger the appropriate color without an explicit human command.

6.5.1 Analysis

The effects of connection reinforcement on inter-modal simulation can be seen in Figures 6-14 and 6-15. The first graph shows the weights between the `Center` Property Node and each of the color speech feature nodes with a mostly consistent mapping between the two modality instances (there is some cross-occurrence towards the end with the `''Red''` `Speech` Feature node).

The second graph shows the effects of mostly alternating contingency between the proprioceptive node and the speech feature nodes. Across the experimental sequence, these reinforcements cancel each other out and result in a low simulation factor between these nodes.

6.6 Summary

In this chapter we have presented an implementation of the research architecture set forth in Chapter 4 and initially explored in Chapter 5. We will next evaluate this implementation in a controlled human subject study.

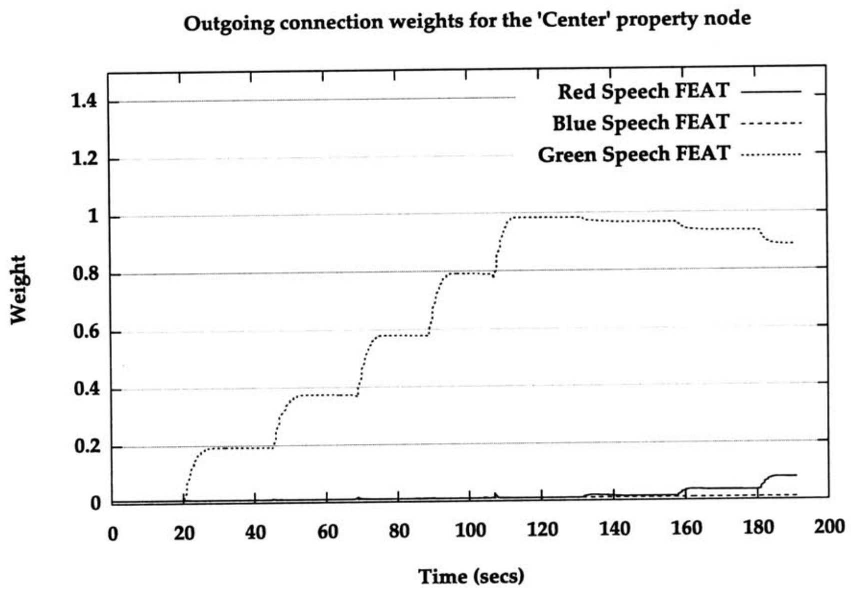


Figure 6-14: Connection reinforcement with consistent contingency. One of the outgoing connection weights from the proprioceptive `Center` property node increases and settles at over 0.9.

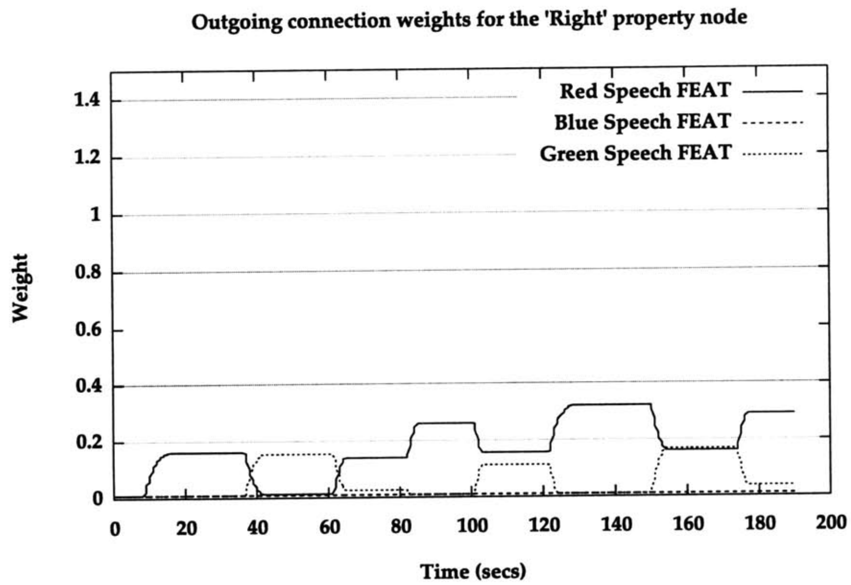


Figure 6-15: Connection reinforcement with inconsistent contingency. None of the outgoing connection weights from the proprioceptive `Right` property node exceeds 0.4.

Chapter 7

Collaborative Lighting Task: Human Subject Study

We have conducted a human subject study to evaluate the performance of the implemented system described in the previous chapter, and the effects it has on the efficiency and fluency of the task, as well as on the human perception of the robot. We were particularly interested in how the system performs within the context of practice, in which the human and the robot repeat a fixed set of identical actions.

7.1 Experimental Design

The study was a between-group controlled experiment with two conditions. The control condition will henceforth be called the `REACTIVE` condition, corresponding to the baseline condition in which both the Emulation/Simulation framework and the Intermodal Reinforcement were disabled. The remainder of the system, i.e. the perceptual network and all activation streams and thresholds, were identically retained. The second condition is denoted `FLUENCY`. In this condition, both the Emulation/Simulation framework and the Intermodal Reinforcement were active with fixed parameters.

To control for instruction bias, neither group was told whether the robot will adapt to their behavior. All participants were allowed to practice with the system before beginning the experiment.

We recruited subjects from the campus and area communities, through email solicitation bearing the following text:

> The Robotic Life Group at the Media Lab is conducting a user study to examine aspects of human-robot collaborative interaction. In this interaction, participants will collaborate on physical tasks with our new robot AUR, a robotic desk lamp. The study will last about 30 minutes, and participants will receive a $10 gift certificate.

A hyperlink to a webpage allowing the recipient to sign up for a slot was also included in the email message.

The experiment included two sequences, or "patterns". Pattern A — indicated by the title "Sequence M" — was associated with a setup in which there were different colors under the different doors. In Pattern B — denoted "Sequence G" for the subjects — there was a one-to-one mapping between location and color, i.e., the same color was hidden under all four doors in a single location. For example, all doors in location B hid the color "Blue", and all doors in location A hid the color "Green". Therefore both the human's memory and the robot's inter-modal reinforcement process could more easily learn the correct association between spatial location and color.

After the subjects were done with both rounds, they were asked to identify the robot as male or female, and asked whether the robot was more of a tool or more of a teammate. Then they were asked to fill out the post-experimental questionnaire.

The experimental protocol was reviewed and approved by the institutional review board of the Massachusetts Institute of Technology.

7.1.1 Subjects

We recruited 38 subject, who were arbitrarily designated to one of the two experimental conditions, and for each subject, they were arbitrarily assigned the order of sequences between Pattern A first, and Pattern B first.

At the last day of the experiment we experienced an unrecoverable hardware failure, forcing us to release the last 5 subjects. We thus remained with 33 subjects, 15 in the REACTIVE condition, and 18 in the FLUENCY condition.

It should be noted that for the first 4 subjects in the FLUENCY condition, the software did not have the correct thresholds (a problem that was later fixed). However, the robot performed appropriately and we retained the data from those subjects.

Two subjects in the FLUENCY condition experienced a mechanical failure. This was resolved in a short period of time, and the subjects continued the experiment. The failure affected the amount of data we were able to obtain from these subjects, a fact we addressed as described in Section 7.4.1.

Subjects were allowed to practice but usually elected not to practice for very long. As a result the first and second sequence attempt of most subjects was significantly slower than other rounds, leading us, for some metrics — for example when considering relative improvement over time — to only consider the second round each subject performed, assuming that at that point the subjects were familiar enough with the system that we measured only trial-related metrics.

7.2 Terminology

In the coming sections we will use the following terminology:

Turn is the time and actions occurring between two consecutive turn buzzers. These include a single event of correctly shining the right light onto the right board.

Sequence is a set of eight turns. There are ten **attempts** at a sequence.

Round is a set of ten attempts at a sequence. The *first round* is the round performed first; similarly for the phrase *second round*. Rounds can also be identifies by *patterns*. In this case we will refer to *Pattern A* and *Pattern B*. Note that for different subjects the order of patterns, i.e. the mapping between patterns and round numbers, is different.

Task is a set of two rounds, using both patterns.

7.3 Results

Building on our experience from the study described in Section 3.4, we have evaluated a number of behavioral hypotheses (**H1–H8**, Section 7.4), as well as a number of hypotheses based on the subjects' self-report (**H9–H17**, Section 7.5).

7.4 Behavioral Measures

The behavioral measures recorded in this experiment were elicited from the log files generated by the experiment software. The software running the robot logged a number of events, including the human's speech commands, the robot's action selection, the human's hand position, and the experimenter's buzzer times.

7.4.1 Data Cleanup

To account for a number of mechanical failures as mentioned above, as well as for mistakenly recorded turn and sequence buzzer events, the data has been automatically cleaned up, by eliminating the following sequences:

- Any sequence attempt that does not contain 7–9 turns was eliminated.

- Any sequence attempt that lasted for less than 25 seconds or more than 180 seconds was eliminated.

As a result, two subjects' data included only 8 sequences in one of the rounds, one subject's data remained with 9 sequences in both rounds, and two subjects' data remained with 9 sequences in one of their two rounds.

We have included the valid data from these subjects in our analyses, except in Hypothesis **H1** below. In **H1**, as well as in the graphs depicting sequence-by-sequence progress on the recorded behavioral measures, we included only data from trials containing 10 valid sequences.

Figure 7-1 shows turn and sequence data from a subject experiencing a mechanical failure before and after cleanup.

7.4.2 Team Performance

Our first set of hypotheses were concerned with the performance of the human-robot team. The following metrics were elicited from the log files recorded by the experiment software:

`TASK` — The overall time to complete all ten attempts of a single sequence pattern.

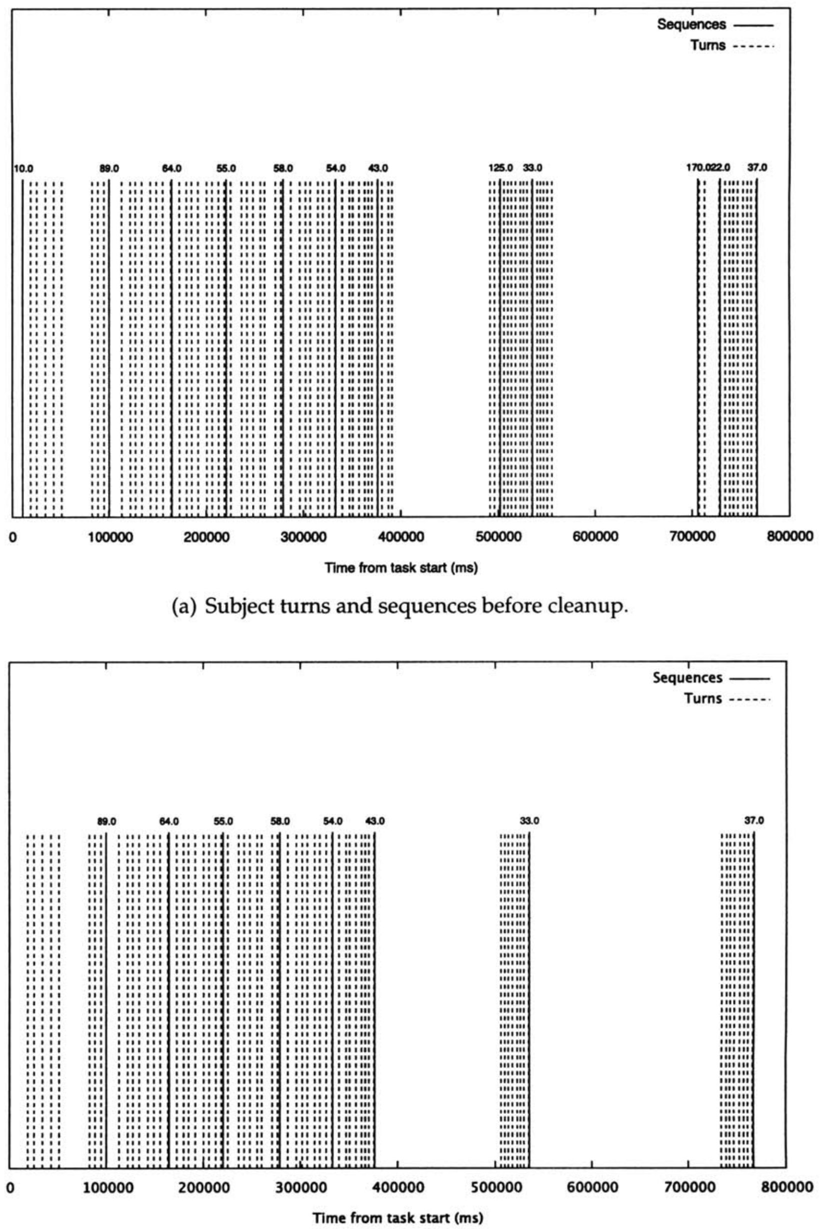
`SEQ` — The time to complete a single eight-turn sequence.

`TURN` — The time to complete a single turn within a sequence.

The metrics are indexed as follows:

$$\mathtt{METRIC}_{round}^{seq\_attempt}(turn)$$

Where *round* can be indicated by a number 0 or 1 meaning the *first* or *second* round; or a letter *A* or *B* indicating which pattern this round corresponds to.



(a) Subject turns and sequences before cleanup.

(b) Subject turns and sequences after cleanup.

Figure 7-1: Effects of cleanup on one round of one subject, who experienced a mechanical failure of the robot during their seventh and ninth sequence attempt. Sequence attempts 1–6, 8, and 10 were retained for data analysis.

Using the above notation, $\texttt{SEQ}_0^4$ indicates the time it took a subject to complete the 5th sequence attempt in the first round.

We tested the following hypotheses:

H1 — Overall task completion time

$\texttt{TASK}_0 + \texttt{TASK}_1$ is significantly lower in the FLUENCY condition compared to the REACTIVE condition.

H2 — Mean sequence attempt time

$mean(\texttt{SEQ}_0) + mean(\texttt{SEQ}_1)$ is significantly lower in the FLUENCY condition compared to the REACTIVE condition.

H3 — Mean sequence attempt time (last 5 sequence attempts)

$mean(\texttt{SEQ}_0^{5-9}) + mean(\texttt{SEQ}_1^{5-9})$ is significantly lower in the FLUENCY condition compared to the REACTIVE condition.

H4 — Best sequence attempt time

$min(\texttt{SEQ}_0) + min(\texttt{SEQ}_1)$ is significantly lower in the FLUENCY condition compared to the REACTIVE condition.

H5 — Improvement in second round sequence time (last over first attempt)

$\texttt{SEQ}_1^9/\texttt{SEQ}_1^0$ is significantly lower in the FLUENCY condition compared to the REACTIVE condition.

Note: We use only data from the second round under the assumption that, at this point, the subject is familiar with the task structure and robot, and we thus measure only the team's improvement and not the initial practice needed by the subject.

Table 7.1 shows the values measured for the above performance metrics and the statistical significance of the difference between the two experimental conditions. All significance evaluations in this chapter were performed using a T-test with independent samples.

Table 7.1: Human-robot team performance metrics. Values are *mean* ± *s.d.*. in seconds, except in H5, which is a fraction.

Hypothesis	Metric	REACTIVE	FLUENCY	T	p	
H1 — Total task time	$\mathtt{TASK}_0 + \mathtt{TASK}_1$	1401.66 ± 162.90	1196.26 ± 226.83	t(24)=2.609	$p < 0.05$	*
H2 — Mean seq. time	$mean(\mathtt{SEQ}_0) + mean(\mathtt{SEQ}_1)$	141.38 ± 17.38	116.21 ± 22.67	t(30)=3.487	$p < 0.01$	**
H3 — Mean seq. time (2nd half)	$mean(\mathtt{SEQ}_0^{5-9}) + mean(\mathtt{SEQ}_1^{5-9})$	131.04 ± 17.76	97.93 ± 22.75	t(30)=4.544	$p < 0.001$	***
H4 — Best seq. time	$min(\mathtt{SEQ}_0) + min(\mathtt{SEQ}_1)$	117.93 ± 16.11	83.87 ± 18.81	t(30)=5.461	$p < 0.001$	***
H5 — Seq. time improvement	$\mathtt{SEQ}_1^9 / \mathtt{SEQ}_1^0$	0.76 ± 0.18	0.50 ± 0.15	t(30)=4.718	$p < 0.001$	***

Table 7.2: Human-robot team fluency metrics. Values are *mean* ± *s.d.*. H7, H8 is measured in seconds

Hypothesis	Metric	REACTIVE	FLUENCY	T	p	
H6 — Human idle time	$mean(\mathtt{IDLE}_0, \mathtt{IDLE}_1)$	0.46 ± 0.08	0.364 ± 0.09	t(30)=3.001	$p < 0.01$	**
H7 — Robot funct. delay	$mean(\mathtt{DELAY}_0, \mathtt{DELAY}_1)$	4.81 ± 9.91	3.66 ± 15.72	t(30)=2.434	$p < 0.05$	*
H8 — Robot funct. delay (2nd half)	$mean(\mathtt{DELAY}_0^{5-9}, \mathtt{DELAY}_1^{5-9})$	4.07 ± 10.97	1.48 ± 6.14	t(30)=3.487	$p < 0.001$	***

All of our hypotheses were confirmed, demonstrating a significant improvement in team performance under the FLUENCY condition. Note that examining in particular the second half of each task round (as in hypotheses **H3**, **H4**, and **H5**) leads to an increase in difference between the two conditions, and an increase in significance. This phenomenon is further illustrated in the figures below.

Figures 7-2 and 7-3 show the average sequence attempt time (SEQ) for both conditions, split by round. The notion of initial practice runs is evident in this figure, as the second round starts at a lower time than the first round. In both cases, the FLUENCY condition converges at slightly over 40 seconds, while the REACTIVE condition maintains an average over 60 seconds.

Figure 7-3 shows data for sequence pattern *A* and *B*, respectively. Since the robot and the human "learn" the color sequences more fully in pattern *B*, we see a more dramatic improvement in the FLUENCY condition, converging on a below-40 second score in the final sequence attempt.

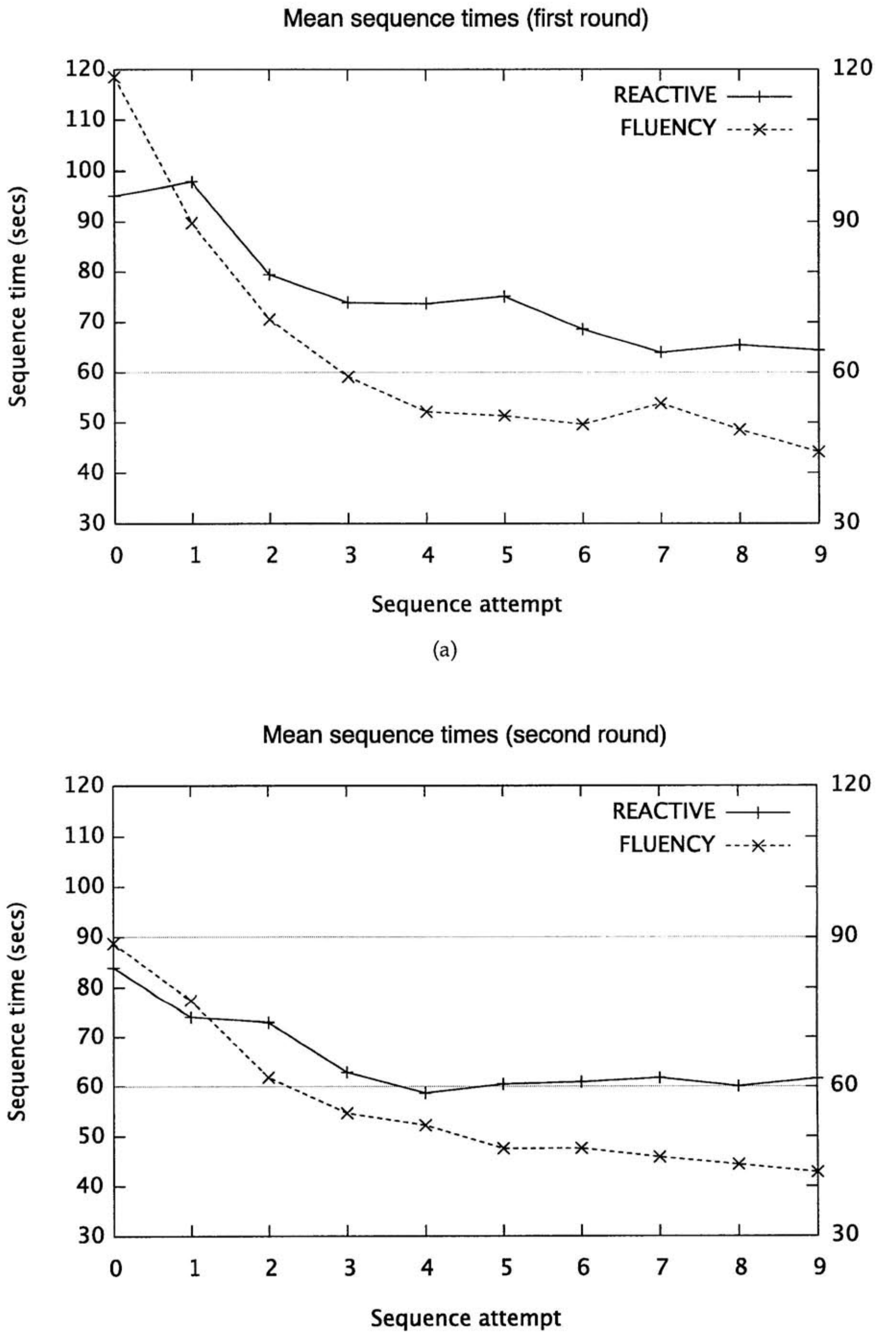
7.4.3 Fluency Metrics

The second set of hypotheses tested in this experiment relate to the fluency of the team. These are based on the fluency metrics set forth in Hoffman and Breazeal [2007] as well as in Chapter 3. The following metrics were elicited from the log files recorded by the experiment software:

IDLE — Human idle time. The percentage of time within each task round in which the human hand was stationary or move very little.

This metric is elicited from the human's hand position. A low-passed sensor with hysteresis is triggered every time the frame-by-frame distance of the hand position crosses a certain threshold.

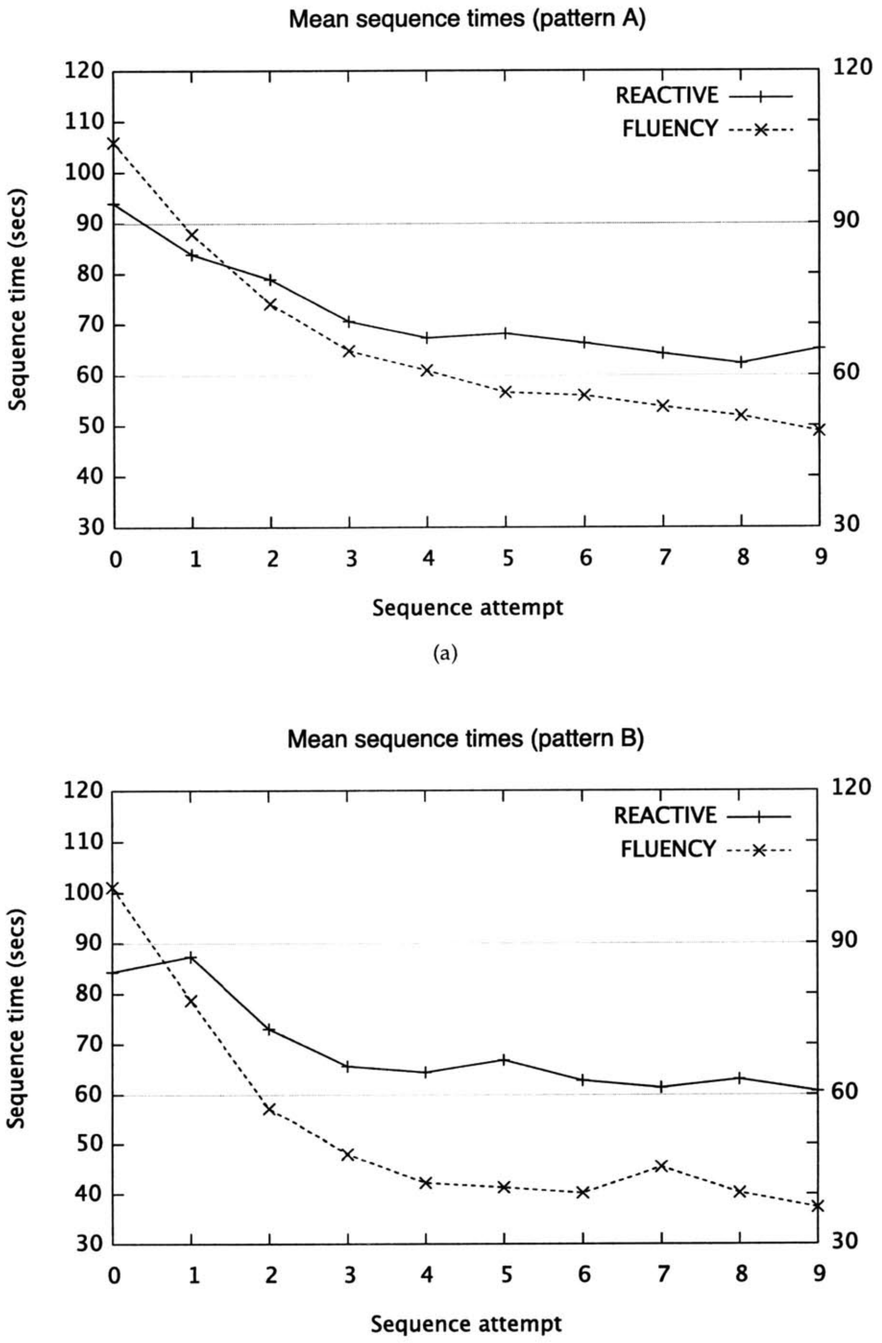
DELAY — Robot functional delay. The time that passed from the beginning of a turn to the onset of the robot's movement.



(a)

(b)

Figure 7-2: Mean sequence times per round.



(a)

(b)

Figure 7-3: Mean sequence times per pattern.

The metrics are indexed as described in the previous section.

We tested the following hypotheses:

H6 — Human idle time

$mean(\texttt{IDLE}_0, \texttt{IDLE}_1)$ is significantly lower in the FLUENCY condition compared to the REACTIVE condition.

H7 — Robot functional delay

$mean(\texttt{DELAY}_0, \texttt{DELAY}_1)$ is significantly lower in the FLUENCY condition compared to the REACTIVE condition.

H8 — Robot functional delay (second half)

$mean(\texttt{DELAY}_0^{5-9}, \texttt{DELAY}_1)^{5-9}$ is significantly lower in the FLUENCY condition compared to the REACTIVE condition.

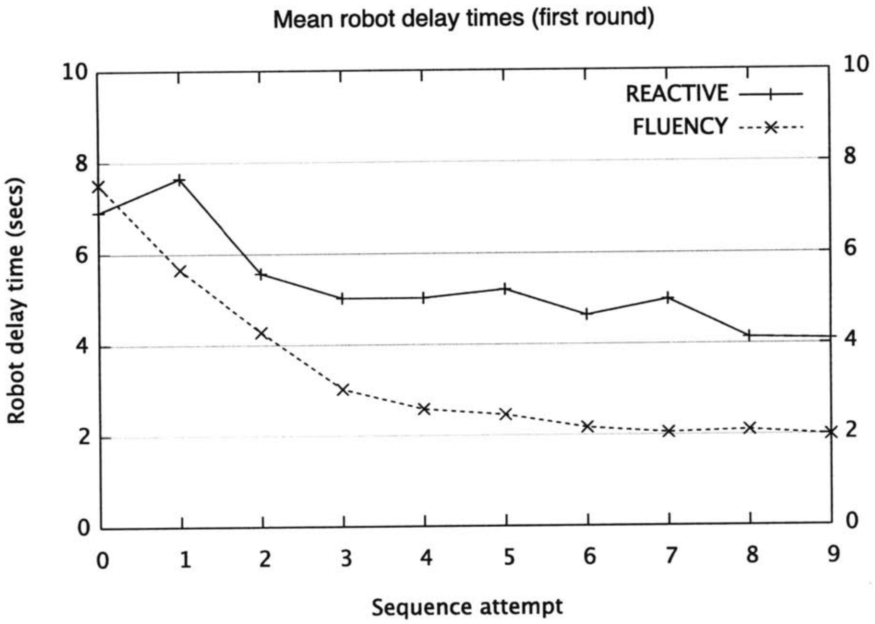
Table 7.2 on page 126 shows the values measured for the above fluency metrics and the statistical significance of the difference between the two experimental conditions.

All of our hypotheses were confirmed, demonstrating a significant improvement in both fluency metrics under the FLUENCY condition. Again, examining the second half of each task round (hypothesis **H8**) shows an increase in difference between the two conditions, and an increase in significance.

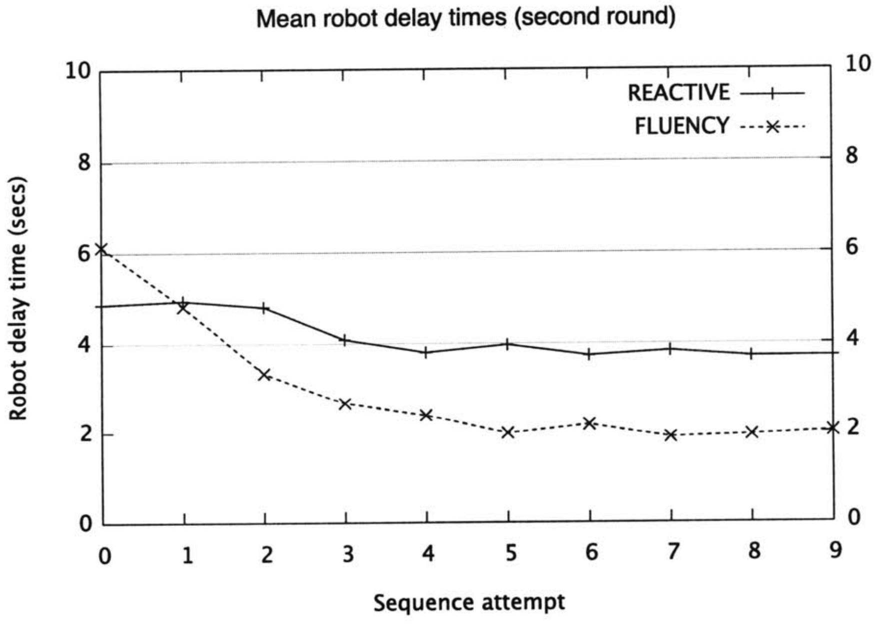
Figures 7-4 and 7-5 shows the average change in robot functional delay time (DELAY) for both conditions, split by trial. It can be seen, in Figure 7-4 (b), that in the REACTIVE condition, the human teammate can do little to improve the robot's delay, after the initial "practice" period.

7.4.4 Relative contribution of human and robot

It is interesting to estimate the relative contribution of each team member to the improvement of the team, and especially to compare the learning rate of the human and the robot as it manifests itself on the temporal change in the team member's contribution.

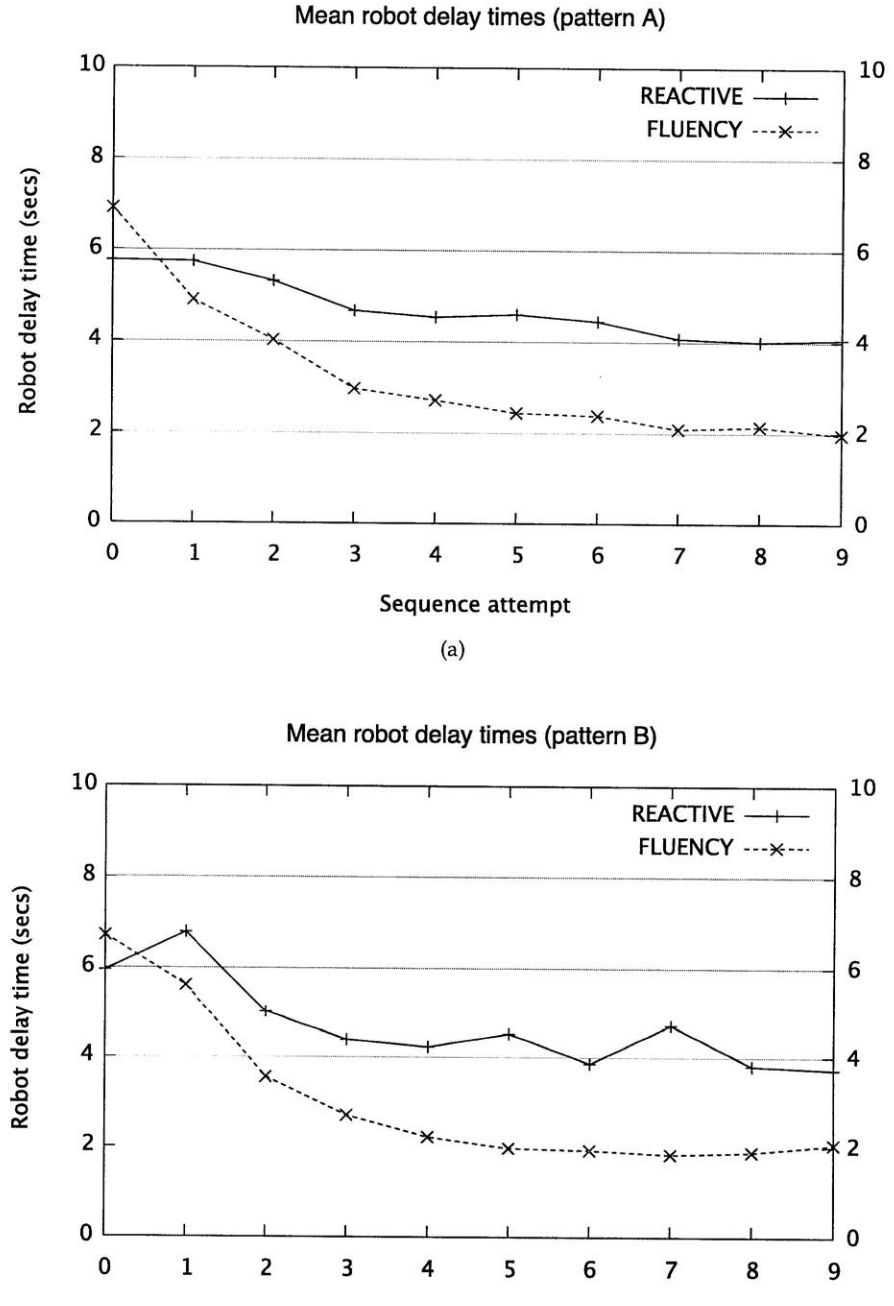


(a)



(b)

Figure 7-4: Mean delay times per round.



(a)

(b)

Figure 7-5: Mean delay times per pattern.

We estimated this measure as follows: For each sequence attempt, we compare a given metric to the value of the same metric in the first attempt. Since the first round included a few practice attempts, we only estimate this measure on the second round for each subject.

As the robot does not adapt or learn in the REACTIVE condition, we consider the improvement of the team in that group to be solely on behalf of the human. We call this "the human contribution" to the team's improvement. Subtracting the human contribution function from the improvement of the team in the FLUENCY condition, we obtain "the robot contribution" to the team's improvement.

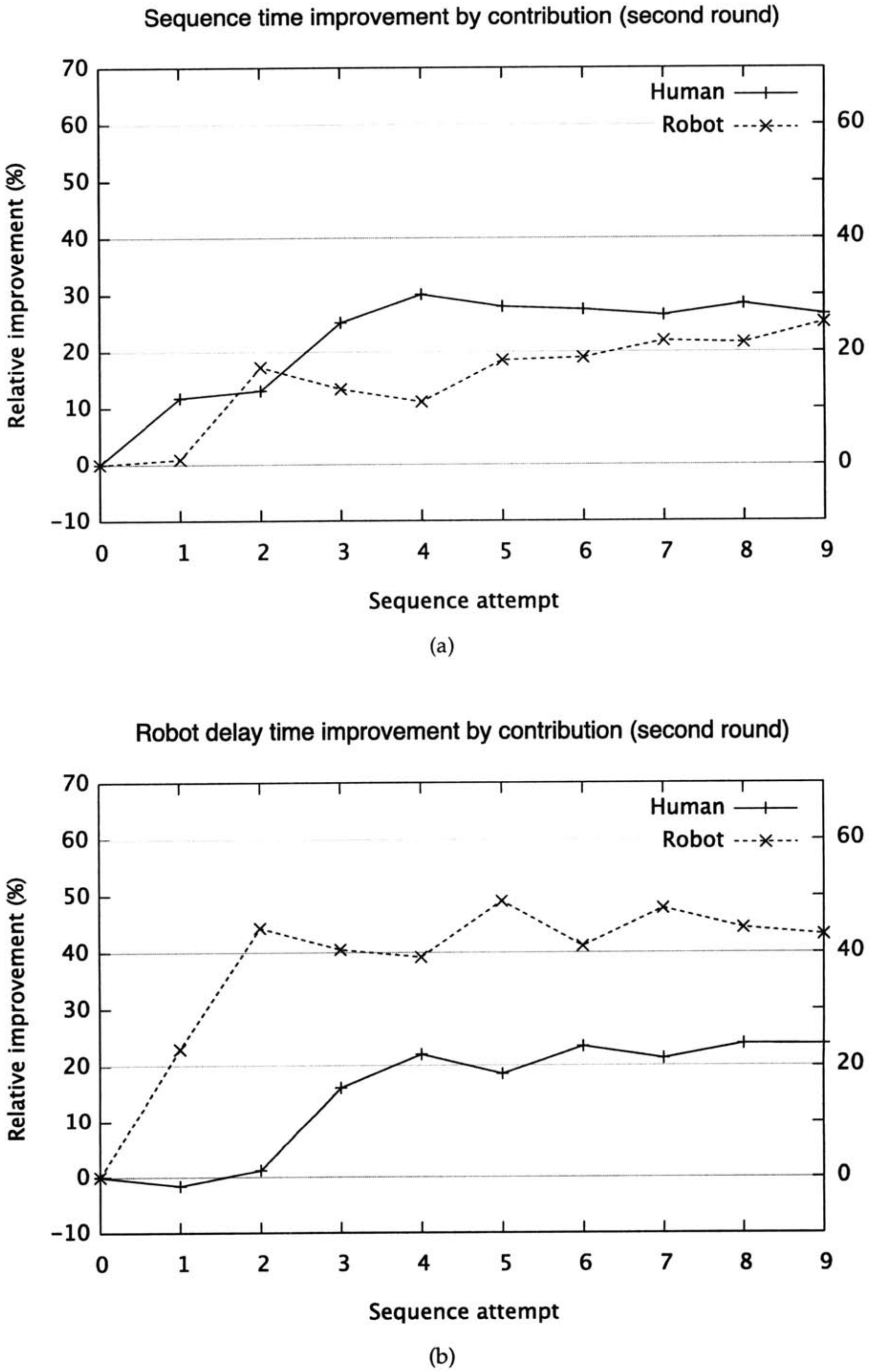
Figure 7-6 shows the relative contribution of the team members on the improvement in sequence time. Note, in (a), that the deduced rate of learning on the robot's part roughly matches the adaptation of the human, which could possibly lead to an increased sense of partnership and "like-me" perception. It is also worth noting, in (b) that the robot's contribution to a lowered DELAY metric (a measure that might be related to fluency) converges on roughly twice the contribution of the human to that metric.

7.5 Self-report questionnaire

In addition to the behavioral metrics we have administered a self-report questionnaire including 41 questions. 38 questions asked the subjects to rank agreement with a sentence on a 7-point Likert scale, with the endpoints and midpoint labeled "Strongly Disagree" (1), "Neutral" (4), and "Strongly agree" (7). Three questions were open ended responses. Please refer to Appendix A for a complete list of questions.

We have compounded the questions into nine scales we propose as valuable to evaluate human-robot teamwork. In these scales, the number refers to the question number on the questionnaire as listed in Appendix Section A.1, whereas *pos*() refers to the value 1–7 on the Likert scale, and *neg*() refers to 8 minus the value on the Likert scale.

We have verified the reliability of these scales within our subject population using Cronbach's alpha measure.



(a)

(b)

Figure 7-6: Relative contribution of the team members on (a) sequence time, and (b) robot delay.

The proposed scales are as follows:

HRT-ENJOY — The overall enjoyment of the teamwork experience

Metric: $[pos(1) + neg(2) + pos(3)]/3$

1 : My overall experience was enjoyable.

2 : My overall experience was boring.

3 : I would like to repeat the task again.

Cronbach's alpha: 0.716

HRT-FLUENCY — The sense of fluency in the teamwork

Metric: $[pos(7) + pos(8) + pos(12)]/3$

7 : The human-robot team worked fluently together

8 : The human-robot team's fluency improved over time.

12 : The robot contributed to the fluency of the interaction.

Cronbach's alpha: 0.801

HRT-IMPROVE — The team improvement over time

Metric: $[pos(6) + pos(8) + pos(11)]/3$

6 : The human-robot team improved over time.

8 : The human-robot team's fluency improved over time.

11 : The robot's performance improved over time.

Cronbach's alpha: 0.793

HRT-ROBOT-CONTRIB — The robot's contribution to the team

Metric: $[neg(22) + pos(23) + neg(24) + pos(25)]/4$

22 : I had to carry the weight to make the human-robot team better.

23 : The robot contributed equally to the team performance.

24 : I was the most important team member on the team.

25 : The robot was the most important team member on the team.

Cronbach's alpha: 0.785

HRT-ROBOT-TRUST — The human's trust in the robot

Metric: $[pos(14) + pos(16)]/2$

14 : I trusted the robot to do the right thing at the right time.

16 : The robot was trustworthy.

Cronbach's alpha: 0.772

HRT-ROBOT-CHAR — The robot's positive character traits

Metric: $[pos(15) + pos(16) + pos(17)]/3$

15 : The robot was intelligent.

16 : The robot was trustworthy.

17 : The robot was committed to the task.

Cronbach's alpha: 0.827

HRT-WAI-BOND — The HRT Working Alliance bond subscale

Metric: $[neg(30) + pos(31) + pos(34) + pos(36) + pos(37) + pos(39) + pos(40)]/7$

30 : I feel uncomfortable with the robot.

31 : The robot and I understand each other.

34 : I believe the robot likes me.

36 : The robot and I respect each other.

37 : I am confident in the robot's ability to help me.

39 : I feel that the robot appreciates me.

40 : The robot and I trust each other.

Cronbach's alpha: 0.808

HRT-WAI-GOAL — The HRT Working Alliance goal subscale

Metric: $[pos(32) + neg(35) + pos(38)]/3$

32 : The robot perceives accurately what my goals are.

35 : The robot does not understand what I am trying to accomplish.

38 : The robot and I are working towards mutually agreed upon goals.

Cronbach's alpha: 0.794

HRT-WAI— The overall HRT Working Alliance

Metric: [HRT-WAI-BOND × 7 + HRT-WAI-GOAL × 3 + *neg*(33)]/12

33 : I find what I am doing with the robot confusing.

Cronbach's alpha: 0.843

The last three scales are adapted from the Working Alliance Inventory [Horvath and Greenberg, 1989], a standard instrument evaluating clinician-patient relationship, to fit a human-robot joint task. We did not include a task subscale as the questions in the original WAI task subscale were very specific to a clinician-patient scenario. We denote this instrument HRT-WAI.

We hypothesized there to be a significant difference in these metrics between the two experimental conditions, specifically that these metrics be higher for the FLUENCY condition.

In addition we hypothesized that the following individual questions will be higher rated in the FLUENCY condition compared to the REACTIVE condition:

HRT-ROBOT-FLUENCY — The robot's contribution to the fluency

12: The robot contributed to the fluency of the interaction.

HRT-HUMAN-COMMIT — The human's commitment to the team

27 : I was committed to the success of the team.

HRT-ROBOT-ADAPT — The robot's adaptation to the human

21 : The robot learned to adapt its actions to mine.

7.5.1 Questionnaire Results

Table 7.3 shows the results for the questionnaire hypotheses. Figures 7-7 to 7-12 show the results for those scales in which we found significant difference between the two experimental conditions.

Table 7.3: Survey questionnaire results metrics. Values are *mean* ± *s.d.*. on a 7-point Likert scale, with the endpoints and midpoint labeled "Strongly Disagree" (1), "Neutral" (4), and "Strongly agree" (7). All comparisons were made using a T-test for independent samples.

Hyp.	Metric HRT-	REACTIVE	FLUENCY	T	p	
H9	ENJOY	5.133 ± 1.24	5.352 ± 1.14	t(31)=0.528	not signif.	
H10	FLUENCY	4.978 ± 0.96	5.926 ± 0.98	t(31)=2.798	$p < 0.01$	**
H11	IMPROVE	5.156 ± 0.96	6.167 ± 1.09	t(31)=2.797	$p < 0.01$	**
H12	ROBOT-CONTRIB	2.85 ± 1.11	4.0 ± 1.32	t(31)=2.687	$p < 0.05$	*
H13	ROBOT-TRUST	4.9 ± 1.25	5.417 ± 1.28	t(31)=1.167	not signif.	
H14	ROBOT-CHAR	4.8 ± 1.32	5.407 ± 1.17	t(31)=1.248	not signif.	
H14	WAI-BOND	4.171 ± 1.05	4.365 ± 1.03	t(31)=0.533	not signif.	
H14	WAI-GOAL	3.644 ± 1.47	4.704 ± 1.31	t(31)=1.248	$p < 0.05$	*
H14	WAI	4.139 ± 0.94	4.54 ± 0.94	t(31)=1.248	not signif.	
H15	ROBOT-FLUENCY	4.733 ± 1.22	6.111 ± 1.18	t(31)=3.282	$p < 0.01$	**
H16	HUMAN-COMMIT	6.4 ± 0.74	5.833 ± 1.10	t(31)=1.702	not signif.	
H17	ROBOT-ADAPT	3.467 ± 1.46	5.944 ± 1.06	t(31)=5.656	$p < 0.001$	***

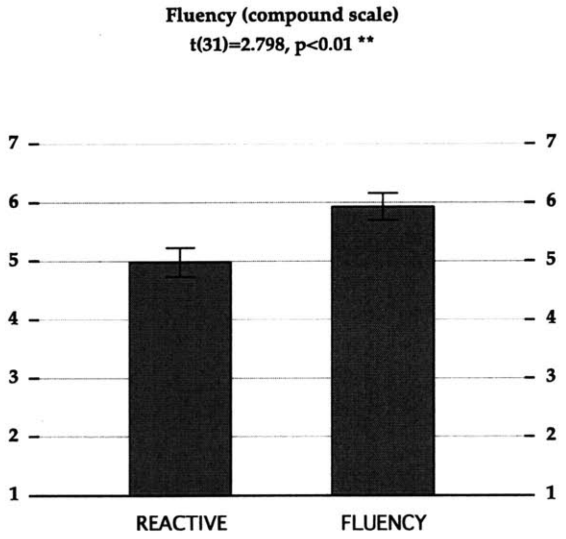


Figure 7-7: Significant differences on the HRT-FLUENCY scale between the FLUENCY and the REACTIVE conditions. Error bars indicate standard error.

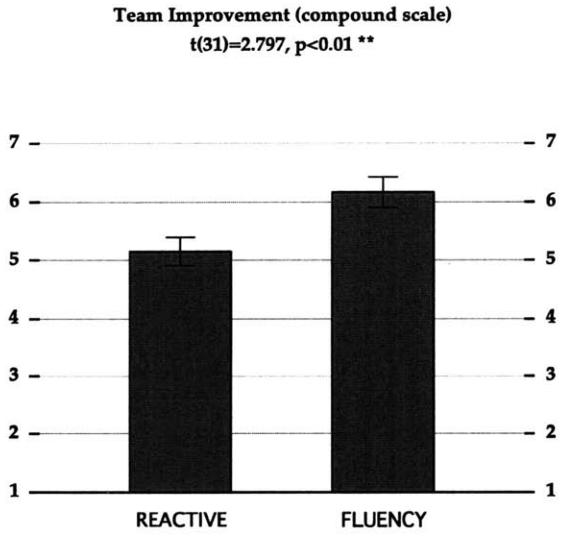


Figure 7-8: Significant differences on the HRT-IMPROVE scale between the FLUENCY and the REACTIVE conditions. Error bars indicate standard error.

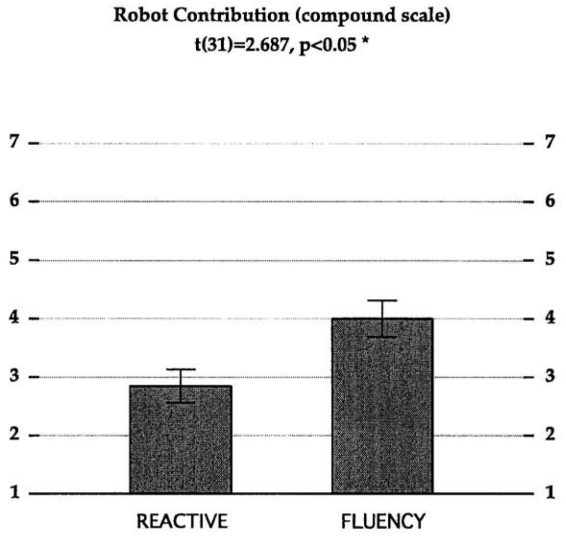


Figure 7-9: Significant differences on the `HRT-ROBOT-CONTRIB` scale between the `FLUENCY` and the `REACTIVE` conditions. Error bars indicate standard error.

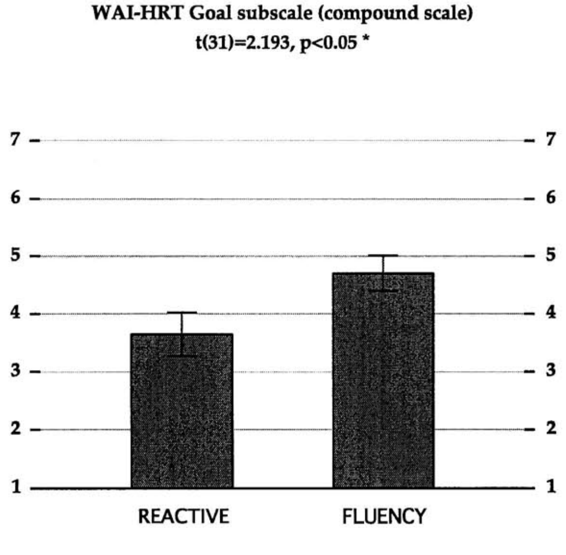


Figure 7-10: Significant differences on the `HRT-WAI-GOAL` scale between the `FLUENCY` and the `REACTIVE` conditions. Error bars indicate standard error.

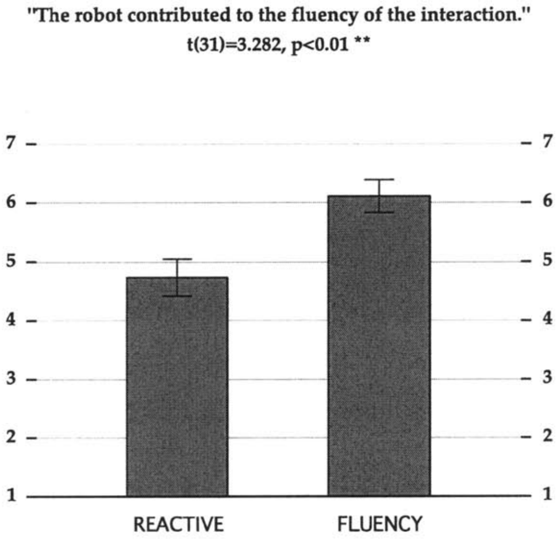


Figure 7-11: Significant differences on the HRT-ROBOT-FLUENCY scale between the FLUENCY and the REACTIVE conditions. Error bars indicate standard error.

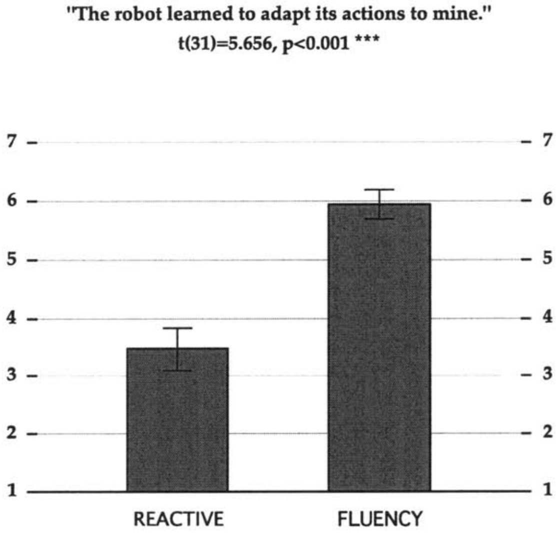


Figure 7-12: Significant differences on the HRT-ROBOT-ADAPT scale between the FLUENCY and the REACTIVE conditions. Error bars indicate standard error.

Table 7.3 reveals significant differences between subjects in the two experimental conditions with regard to the fluency scales in the questionnaire. Both the HRT-FLUENCY and the HRT-ROBOT-FLUENCY measures are significantly different at $p < 0.01$.

Additionally, subjects in the FLUENCY condition rated the robot's contribution to the team significantly higher than subjects in the REACTIVE condition, as well as the team's over all improvement.

While these task-related scales differ significantly, we were not able to show a significant difference in the robot's perceived positive character traits (intelligence, trustworthiness, and commitment), in the trust the human put in the robot, in the human's commitment to the task — which was incidentally higher for the REACTIVE condition, if not significantly so — or in the subject's overall enjoyment of the experiment. While all but one of these scales were higher for the FLUENCY condition, these differences were not statistically significant.

We believe that this is in part due to the low expectation people have of robots, which caused the evaluation of the reactive robot to be high as a response to the robot's generally reliable functioning. This hypothesis could be evaluated in a within-subject experiment comparing the two robot architectures.

Note that while the overall robot's character was not rated significantly different between the two conditions, the robot's intelligence was (see Appendix Section A.2).

On the HRT-WAI scale, the goal subscale was significantly different between the two conditions, while the bond subscale — as well as the overall HRT-WAI score — were not. One possible explanation for that phenomenon could be that it takes longer than the experiment's duration to form a bond, whereas the mutual agreement on goals can be established in a shorter time span.

Gender differences

Appendix Section A.3 lists gender differences at $p < 0.05$ on survey questions. The most significant difference, at $p < 0.01$, was on question 24 ("I was the most important team member on the team"), shown in Figure 7-13. As a result, there was a significant difference in the HRT-ROBOT-CONTRIB scale, shown in Figure 7-14.

We found this to be mostly due to the difference in the FLUENCY condition (Figures 7-15 and 7-16).

In the REACTIVE condition there was no gender difference on this scale (see: Figures 7-17 and 7-18). Similarly, there was no significant difference on question 24 in the REACTIVE condition, although female subjects did rate their importance lower than male subjects (see: Figures 7-17 and 7-18).

All comparisons were made using a T-test with independent samples.

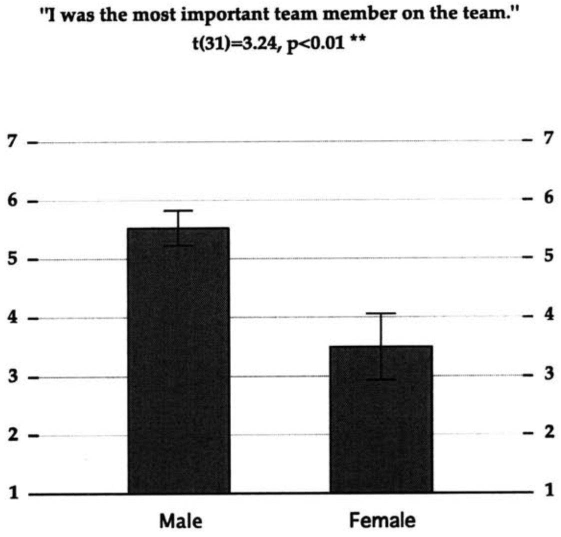


Figure 7-13: Gender differences on Question 24 between male and female subjects. Error bars indicate standard error.

7.5.2 Oral questions

Subjects were also asked orally, whether they considered the robot more of a tool or more of a teammate. While the answers were biased towards the teammate in the FLUENCY con-

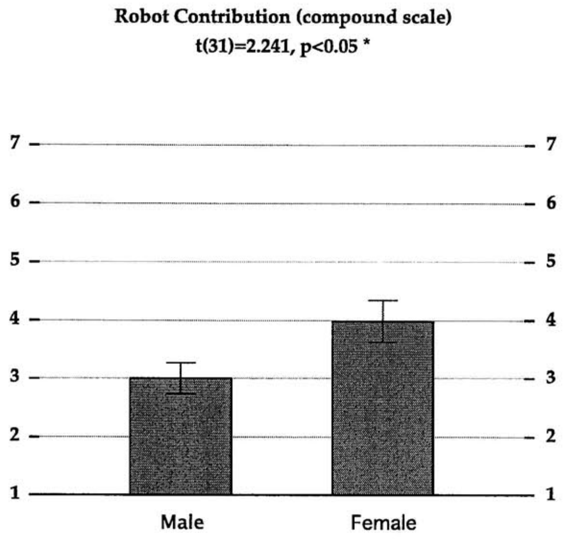


Figure 7-14: Gender differences on the HRT-ROBOT-CONTRIB scale between male and female subjects. Error bars indicate standard error.

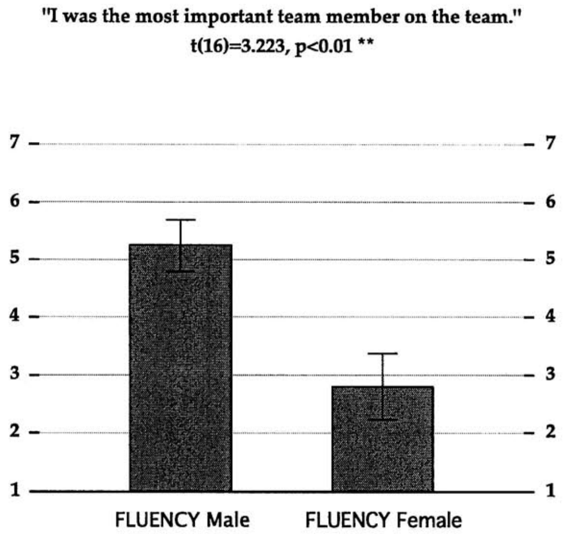


Figure 7-15: Gender differences on the Question 24 between male and female subjects in the FLUENCY condition. Error bars indicate standard error.

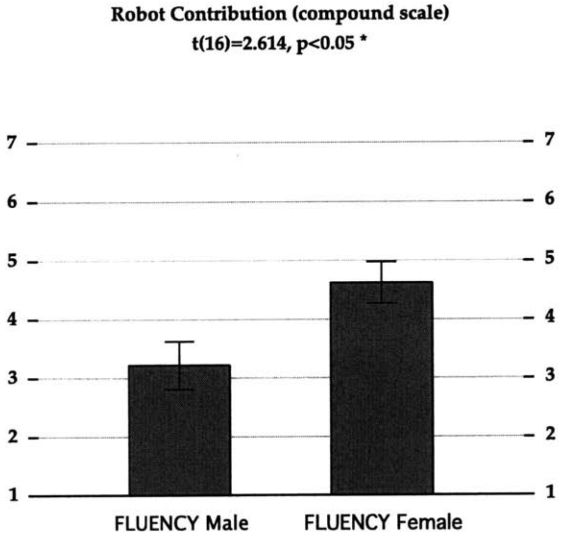


Figure 7-16: Gender differences on the HRT-ROBOT-CONTRIB scale between male and female subjects in the FLUENCY condition. Error bars indicate standard error.

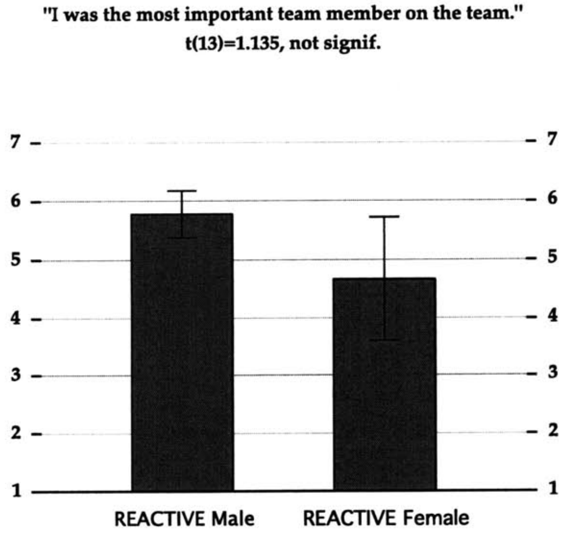


Figure 7-17: Gender differences on the Question 24 between male and female subjects in the REACTIVE condition. Error bars indicate standard error.

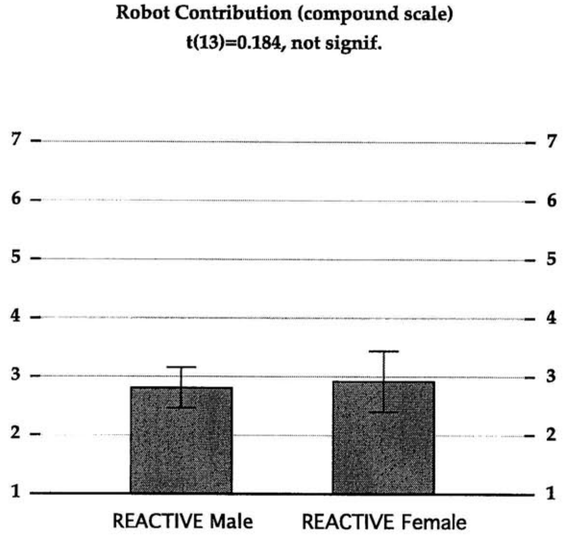


Figure 7-18: Gender differences on the HRT-ROBOT-CONTRIB scale between male and female subjects in the REACTIVE condition. Error bars indicate standard error.

dition, we could not find a statistically significant difference. Table 7.4 shows the responses to these questions by condition.

Condition	Teammate	Tool
FLUENCY	8	4
REACTIVE	4	7

Table 7.4: Responses, by condition, to the question whether the robot acted more as a tool or more as a teammate.

To the question whether the robot was male or female, all but one subjects chose "male".

7.5.3 Open-ended responses

Anecdotal evidence from the open-ended response section of the questionnaire reveals several differences between the two conditions. Overall, the qualitative response of subjects in the FLUENCY condition seems to be more favorable than that of subjects in the REACTIVE condition.

Positive comments in the FLUENCY condition included subjects reporting to be "highly impressed at [the robot's] learning and movement speed, at [its] capacity to retain information and elaborate upon it", and a subject saying that the robot "worked well, and I felt a sense of relief/relaxation when it just did what I was about to tell it to do." One subject reported to have "had emotional responses that went from tenderness (a lesser entity than myself) to amusement to respect (it has better memory!) and trust." And one went so far as to claim that "[b]y the end of the second sequence, we were good friends and high-fived mentally after the task was done."

Such positive comments were rare in the REACTIVE condition. Characteristic descriptions of the robot in that condition were "It did exactly what is was supposed to", "The robot performed fairly well, but it did not understand the larger plan of action", and "[T]he lamp was not creepy". One subject, however, remarked that "[t]here were a few moments when I felt like I was ineracting with a being that was more alive... [than a] machine".

Several negative comments, in particular with regards to the robot's contribution as a team member, were peppered throughout the comments of subjects in the REACTIVE condition. These included "The robot was more of an assistance than an active team member", "If the plan could be interpreted from speech, even if it was a simple plan, I would have felt like the robot was even more of a partner", "I felt like I was controlling the robot, rather than it being part of a team. I didn't really feel like there was any symbiotic interaction between myself and the robot; rather, it feel like a tool and something that I command, rather than an equal partner", and "It did pretty well but I was definitely driving it, and it just fell like a lazy apprentice". This also reflected on the overall sense of the team's accomplishment, in remarks such as "i'm not sure our team performance ever improved."

In contrast, subjects in the FLUENCY condition remarked on the robot's contribution to the team, and referred to it several times as a teammate: "By the end of the first sequence I realized that he could learn and work as my teammate", "my interaction with the robot was not that different than with a human teammate. i sometimes believed him (!) [*sic*] IT a better performer than myself and was impresed at the rate of improvement we had", and "i would love having a robot as a teammate to perform any task, my fear is seeing how easily and fast i grew affectionate to it and wondering if i'll ever have the need for it to have true (!) [*sic*] feelings for me or discovering i dont mind it they are not human."

One subject in the FLUENCY condition said, however, that "i did not perceive it as human but more as kind of a thing, possibly an animal. i think this might have to do with the fact that i was asked go give it short commands similar to the ones given to animals."

Self-deprecation in the FLUENCY condition

A surprising effect of the experiment was that in the FLUENCY condition there were a high number of self-deprecating comments, and comments indicating worry or stress because of the robot's performance. This was also evident in informal conversation with subjects after the experiment, in which several subjects in that condition remarked on stressful feelings that they weren't performing at an adequate level.

Written remarks included "I would essentially forget the pair of colors I had memerized [*sic*] - this slowed me down on the second sequence", "The robot is better than me","The performance could had been better if I didn't make those mistakes", "[I] worried that I might slow my teammate down with any mistakes I might have made", "Apparently she learned faster than I did", and even "I am obsolete". There were no similar comments in the REACTIVE condition.

While it is beyond the scope of this dissertation to further explore this aspect of our findings, it should raise a warning flag to any designer in the human-robot interaction field. The prevalence of this response may indicate a need for a human to feel in some way more intelligent than the robot they are interacting with.

Maintaining the balance of increased robot responsiveness and the intimidation that might result is generally an overlooked aspect of HRI, which these results urge us to consider.

Lexical analysis

An independent judge analyzed the comments given by the study subjects according to 15 categories. All-in-all there were 54 comments, 24 by subjects in the REACTIVE condition, and 30 by subjects in the FLUENCY condition. These comments were presented to the judge in randomized order and without indication of the condition they belonged to. For each comment, and for each category, the judge was asked to answer whether the comment falls into the category.

Table 7.5 shows the categories judged, and the percentage of comments in each condition rated as matching each category. Figure 7-19 depicts the lexical analysis data graphically.

"The subject's comment..."	REACTIVE	FLUENCY
	%	
"... portrayed the robot in a positive manner."	20.8	56.7
"... portrayed the robot in a negative manner."	50.0	13.3
"... portrayed the robot as one would a human."	16.7	43.4
"... portrayed the robot as one would a creature."	33.3	13.3
"... portrayed the robot as a teammate."	33.3	30.0
"... portrayed the robot as a tool."	29.2	23.3
"... described the robot in emotional terms."	20.8	33.3
"... attributed credit for the task's success to the human."	20.1	10.0
"... attributed credit for the task's success to the robot."	4.2	30.0
"... attributed blame for the task's problems to the human."	12.5	26.7
"... attributed blame for the task's problems to the robot."	50.0	16.7
"... attributed a gender to the robot."	0	13.3
"... described the robot as intelligent."	0	32.1
"... described the robot as unintelligent."	29.2	9.5
"... was self-deprecating."	4.2	30.0

Table 7.5: Lexical analysis of subjects' open-ended comments

This analysis shows several noteworthy differences: in their open-ended questions, sub-

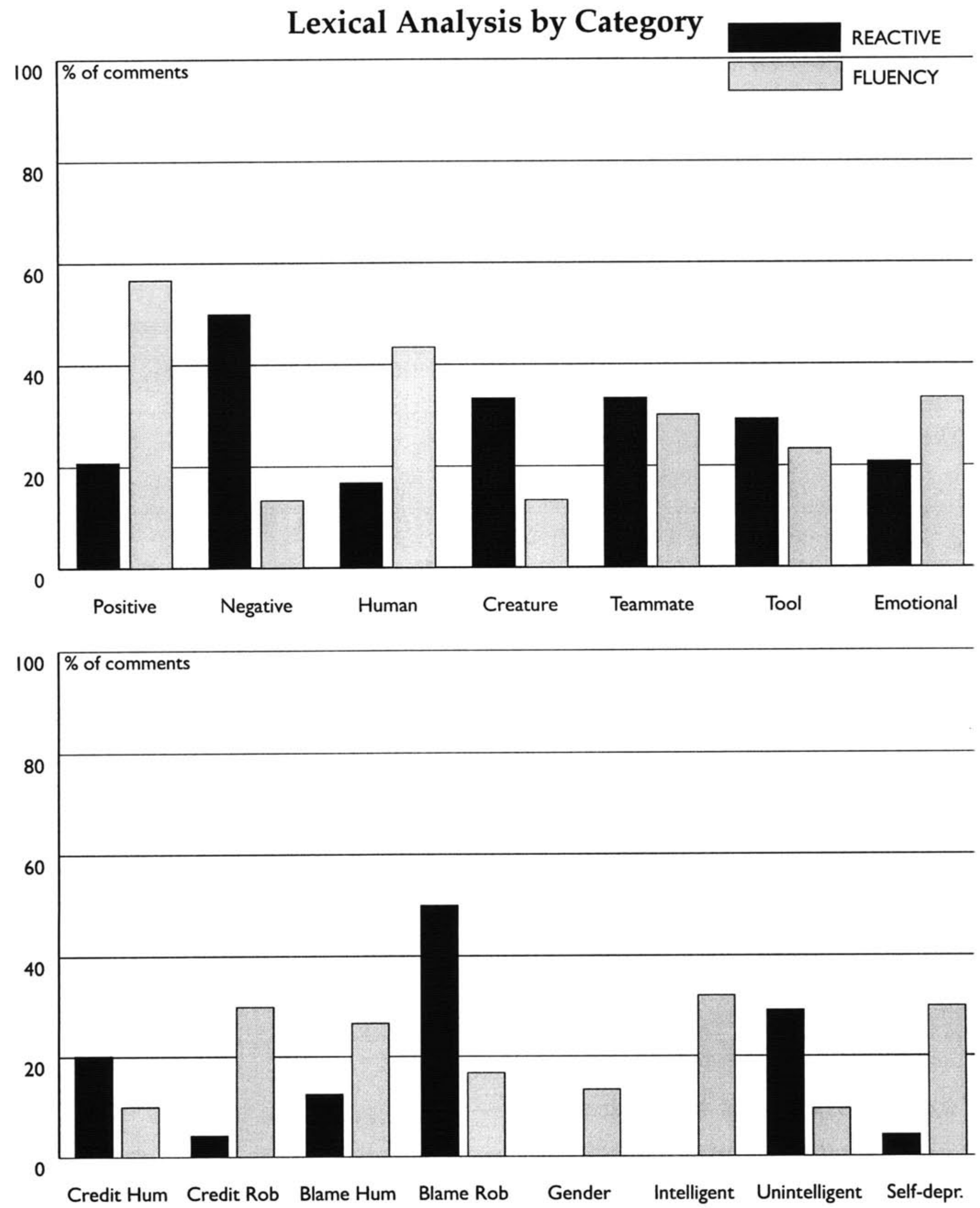


Figure 7-19: Lexical analysis of subjects' open-ended comments (see: Table 7.5)

jects in the FLUENCY condition commented on the robot more positively, and subjects in the REACTIVE condition commented on the robot more negatively. FLUENCY subjects attributed more human characteristics to the robot, although there is little difference in the emotional content of the comments. Similarly, there is no difference in the evaluation of the robot as a tool or a teammate.

Subjects in the FLUENCY condition tended both to give more credit to the robot, and attribute more blame to themselves. Those in the REACTIVE condition far exceeded the other subjects in putting blame on the robot.

Gender attributions, as well as attributions of intelligence only occurred in the FLUENCY condition, while subjects in the REACTIVE conditions tended to comment on the robot as being unintelligent.

Finally, we did find the tendency to self-deprecating comments much more prevalent in the FLUENCY condition.

7.6 Practice revisited

Given the above findings, it makes sense to summarize the notion of practice in our approach. According to the model put forth here, practice operates on two levels: a learning and/or rhythm subsystem triggers top-down simulation biasing perceptual processing and thus altering reaction times and action behavior on the robot's part. Through this mechanism the robot can act based on less perceptual input, given that the same perceptions are simulated in the top-down system. This kind of practice in effect implements the principle that repeated activity shifts action selection from a more robust, but slower decision making process to a more automatic, faster, but less dynamic and reliable process. As expectations become more ingrained and action selection happens based on less perceptual information, the robot is able to act faster, effectively short-circuiting the relation between perception and action.

A second kind of practice is embodied in the connection reinforcement mechanism described in Section 4.7 on page 81. This subsystem alters the weights of connections within the perception and motor activation networks and is another manner in which more "deliberate" action selection procedures (stemming from the motivational layer) transfer to a faster action activation mechanism.

We believe that human practice employs similar mechanisms, and as part of the human-subject study described in this chapter, we often observed similar behavior (such as a "double-take" resulting from a mistaken anticipatory decision) in human and robot as they proceeded to practice the task. It may well be possible that these embodied similarities between the human and the robot team members contributed to a sense of "like-me" perception leading to the humanizing found in the post-experimental survey.

In Section 7.4.4 we showed an empirical comparison of the rate of practice between the human and the robot, finding similarities that are in line with this claim.

In addition, it is known that the perception of an action and its performance are linked in human-human joint action through so-called "mirror neuron" mechanisms (see: Section 2.1). It would be interesting to evaluate how joint practice of the type described in this dissertation affects these mirror mechanisms, and how they interact with human subjects' perception of the robot's human-likeness.

Part III

Framework

Chapter 8

HRI: Four Lessons from Acting Method

As this dissertation draws inspiration from the theories of acting method and practice, we have decided to include a chapter elaborating some core concepts appearing in these practices, which bear relevance to human-robot interaction design. We argue that there are interesting parallels between the challenges actors take on and the ones that designers of HRI robots need to tackle, and therefore lessons to be learned from acting methods for HRI development.[1]

We find this parallel particularly compelling considering the fact that acting method has undergone a significant revolution in the early 20th century, moving from what was essentially a symbol processing approach (for example, the DelSarte system of expression) to an embodied methodology (at the core of the Stanislavski system). This thesis argues that a similar change in perspective is underway in the cognitive sciences, has been gaining momentum in the robotics and artificial intelligence fields, and is worthwhile to be seriously considered by the HRI community.

[1] A version of this chapter was previously published in Hoffman [2005a].

8.1 HRI Design as Acting

Relating the two fields of acting and HRI stems from the observation that actors are trained to be highly tuned to the technical physicality of behavior for various actions and attitudes. An actor's preparation of a role includes a systematic investigation of what gesture, body pose or physical action best describes the internal drive and objective of their character in different contexts. Good actors pay attention to conventions of nonverbal communication and often need to take on the difficult task of portraying complex emotions without words.

These activities are similar to the ones one engages in when designing robotic behavior for human-centric robots. There, too, one is concerned with the readability of expressive behavior, with the breakdown and technical analysis of what it takes to convey a complex inner development, and with the translation of all that into what finally results in a series of motion commands.

More poetically, one might say that both human-centric interactive robot designer and actor are in the business of injecting life into a lifeless object, be it a hierarchy of joints and motors for one, or an arc of dry dialog lines and stage directions for the other. Both robot designer and actor have to analytically break down the complex emergent constellation called "behavior" and reconstruct it in an inherently synthetic, but ultimately meaningful way.

As teachers of acting method have been working to establish principles aimed to help actors prepare an analysis of human behavior with the proclaimed goal of artificially re-creating it in an interactive session, their methods may hold valuable lessons for the progress of HRI within and beyond the realm of entertainment robots. This chapter explores these lessons.

There are, of course, significant differences between human-centric robot behavior and acting. Most notably, as Lee Strasberg said, "the actor need not imitate a human being. The actor is himself a human being and can create out of himself." (In Cole and Chinoy [1970], p.626) This obviously does not hold true for a robot, a fact that can be viewed both a curse and a blessing. On the one hand, much of acting technique draws on the personal experiences of the actor, experience that no robot can bring to the table. On the other hand, a significant portion of an actor's most challenging training is purposed to enable her to "let go" of ingrained experience, of her own personality, and allow the fictitious character to enter into that void. In that respect, robots are fortunate to not have a real ego to battle in an attempt to make room for their alter — or stage — ego.

This important difference could lead to the conclusion that there can be no robotic "acting" without there first being a robot personality. This also reflects a concern in the Artificial Intelligence community, predicting that artificial agents will never be intelligent until they are able to accumulate experience and memories over a prolonged period of time.

We agree that in order to seriously apply Stanislavskian methodology to robots, vast personal experience must first be acquired. But we do not share the outlook that the connection between robotics and acting is rendered impossible until robots are able to gather enough experience to truthfully perform the "as if" projection that actors are required.

What, then, can acting theory teach us that can be applied towards human-interactive robots today? A survey of acting method training and theory literature, written primarily in the last few decades, demarcates several themes that reoccur in almost every text and technique, no matter how different from each other (and how much each method claims to revolutionize its predecessor's approach).

These themes will be referred to herein as (1) *psycho-physical unity*; (2) *mutual responsiveness*; and (3) *objective and inner monologue.*

In the following sections we will briefly show how these themes have been addressed by various acting teachers and theorists, and why we think they may provide lessons for HRI designers.

8.2 Psycho-physical Unity

In his book *Games for Actors and Non-actors*, Augusto Boal states:

> "[...] the human being is a unity, an indivisible whole. Scientists have demonstrated that one's physical and psychic apparatuses are completely inseperable. Stanislavski's work on physical actions also tends to the same conclusion, i.e. that ideas, emotions and sensations are all indissolubly interwoven. A bodily movement 'is' a thought and a thought expresses itself in corporeal form." [Boal, 2002]

As discussed in the first part of this dissertation, mounting evidence shows that it is not only that our mental processes and psychological state affect our corporal functions, but that there exists a strong influence which our physical behavior exerts on our mental state.

Acting theorists have acknowledged that behavior is inherently physically grounded. As early as the late 19th century, acting system reformer François DelSarte, notes the feedback from body to mind:

> "A perfect reproduction of the outer manifestation of some passion, the giving of the outer sign, will cause a reflex feeling within." [Stebbins, 1887]

This interrelationship is only gaining momentum since. Spolin [1999] muses "through physical relationship all life springs", Robert Lewis stresses the physical and sensory side of *affective memory* (In Cole and Chinoy [1970]), Acting teacher Sonia Moore reminds us that "it is a fact that in life the whole complex inner world of a human being, every inner experience, is always expressed physically" [Moore, 1968], and stresses the "how we do it"—the physical aspects of the emotional motivation—as a central part of her Stanislavskian method. Ruth Maleczech uses images, corporal representations, to elicit her performance (In Sonenberg [1996]), as does Boal in his "Image Theater". Broadhurst writes: "[...] technology's most important contribution to art is the enhancement and reconfiguration of an aesthetic creative potential which consists of interacting with and reacting to a physical body, not an abandonment of that body" [Broadhurst, 2004]. And the list goes on. To conclude with Strasberg's more poetically leaning words: "The actor has no piano. In the actor, pianist and piano are the same" (In Cole and Chinoy [1970]).

Following this indication, any work in robotics is well-positioned, exploring the physically grounded intelligence of robots rather than abstract computational processes. The fact that robots *have* bodies is a good starting point, and much of the Aritficial Intelligence community seems to increasingly accept that intelligence is inherently embodied.

However, in HRI robots designed today, cognitive models are still distinctly separate from their physicality. While admittedly immersed in sensory experience, not much attention has been paid to the physical representation of cognition. The "behavior systems" of robots are mostly action selection mechanisms driven by sensory input, and in most HRI implementations the virtual agent driving the robot is an extension of an ideal logical process that communicates with the physical results of its decisions as input and output streams to and from a distant module.

Recent trends in cognitive psychology support what most acting teachers seem to know well: the relationship between mind, decision and corporal function is not one of two distinct systems drawing from each other in a remote-controlled way, but are two aspects of the same; two sides of a Boalian "unity".

The one field of robotics where a similar message seems to be most prominently incorporated is that of close feedback motion control, in which intelligence is employed towards simple, bounded, physical activity—such as balancing a pole or stabilizing a walk. These applications, however, are contained to the most rudimentary activities, and do not scale towards the physical aspects of higher level cognition, which probably should be viewed in the same manner.

In this thesis we argue that HRI design should consider tearing down the implicit barrier between the "motor system", the "perception system", and the "behavior system" and think afresh about a combined architecture where both are sides of the same behavior.

One radical, yet conceivable way forward is to constrain artificial cognitive process to physical embodiments of that process, not allowing any "thinking" that is not physical in nature, simulating more closely our own mento-corporal dependency.

8.3 Mutual responsiveness

New York City acting guru Sanford Meisner is generally associated with the stressing of responsiveness in acting. His rule, embodied among others in the famous "repetition exercise", states

> "Don't do anything unless something happens to make you do it. What you do doesn't depend on you, it depends on the other fellow." [Meisner and Longwell, 1987]

In another place, he commands that "acting isn't chatter, it's responding truthfully to the other person" [Meisner and Longwell, 1987]. This rule, in Meisner's method, is the key to truthful reaction, and to meaningful behavior in the on-stage collaboration between two actors.

Meisner is by no means the only acting theorist to find that what is really occurring in a scene is not happening within any of the actor's heads, or even in their behaviors, but in the space between the two actors. Spolin calls this *communication*:

> "[T]he techniques of the theater are the techniques of communicating. The actuality of the communication is far more important than the method used." [Spolin, 1999]

Similarly, Maleczech speaks of *repercussion*:

> "The other actors are, for me, like the bumpers in a pinball machine. I shoot my pinball, my image, and it goes *tch, tch, tch,* bouncing off those bumpers, each hit having its repercussion. Often the next image will come directly from the response of the other actor." (In [Sonenberg, 1996], p.91)

Moore [1968] reminds us that "as in life, you must evaluate the other person and expected results," as well as "you must coordinate your behavior. [...] Ensemble work means continuous inner and external reaction to each other."

This acting maxim suggests a promising venue for interactive robots. After all, robots, due to their impoverished internal lives and limited experience, tend to be inherently reactive machines. From their very roots, interactive performance robots (as early as James Seawright's Searcher (1966) and Watcher (1966)) "moved and emitted different sound patterns in response to movements and light changes going on around them" [Dixon, 2004]. So it seems that building a robot that can engage in *some* version of the repetition game might not to be an impossible feat.[2]

However, pure mechanical repetition can hardly be considered truly responsive. The missing link in order for repetitive agents to be useful partners in Meisner's repetition game, is solving the undoubtedly harder second part of the game, which is to know when to change the repetitive behavior. For mutual responsiveness is not merely responding in a predictive way, but knowing when to break the mirroring and as a result—switching one's action. It is this failure to break away from the predictive pattern, that causes realistic mutual responsiveness to seem, for the most part, inadequate in HRI robots.

To overcome this limitation we must address the challenge of choosing between simple reactive behavior and breaking away from simple reaction. Detecting the right time and manner to do that is not an easy perceptual processing feat, but may well hold the key to solving the mutual responsiveness problem.

8.4 Objective and Inner Monologue

Perhaps the most significant contribution to modern acting that has been brought on by the Stanislavski system is the elimination of so-called "representational acting". This method

[2] It could be argued that the early interactive computing project *ELIZA* was a version of a repetition-game agent.

of formalizing a range of emotions and associating them with certain bodily and facial actions is often attributed to the DelSarte system of expression, a methodology that has a symbol-processing nature, which unsurprisingly has caught the attention of some robot designers [Bruce et al., 2001].

> "The actor in the French school [...] asks himself 'What must I *choose to do*?' "
> (In [Cole and Chinoy, 1970], p.626; emphasis mine.)

Teachers of the Stanislavski method stress a decidedly different approach, in which the character's motives, objectives, and intentions weigh stronger than the actual line they deliver, allowing the actor to rise beyond simple representation.

Moore [1968] quotes Eugene Vakhtangov saying "A unit in a role or a scene is a step in moving the through-line of actions toward the goal," as opposed to the unit being a single line of dialog or a stage direction.

Her method also encourages the actors to not memorize the lines, but instead focus on analyzing a scene in terms of moving powers, objectives, obstacles, and intentions, leading to choosing actions, creating images and "what if" behaviors out of *affective memory* to truly understand a scene. The scene is then improvised with disregard to the text, and finally re-implemented out of that improvisation on the actual lines of the scene.

Other teachers share this view. Michael Chekhov speaks of Psychological Gestures that draw on the character's "definite desire" in a scene (in Cole and Chinoy [1970], p.519), and Boal, too, stresses that any particular action results from the character's desires, will, and needs.

This then leads to the concept of the *inner monologue*, which is heavily emphasized by Moore's method (and referred to by Meisner as well).

> "[Your inner monologue] is more important than memorizing your lines [...]. The right inner monologue will bring you to your lines, and you may have entirely different intonations." [Moore, 1968]

The actor's inner monologue is based on the analysis of the character and must carry on the whole time the actor is on stage, whether she says something or not. This inner monologue should usually be laid out in detail while preparing for a role, and lends the actor credibility of an internal process while they're on stage, leading up to the lines and thus preventing the lines to be uttered in a void.

Underlying intention structure is also immensely conducive to a reasonably *behaving* robot, intelligence aside. This is particularly true if we attempt to create life-like behavior.

Cracks showing in a robot's lifelike behavior are most often due to the surfacing of too simple — or worse, random — behavior in the action selection system. The robot seems to be behaving unintelligently whenever it displays a lack of clear and readable objectives, desires and intentions, or when there is no reasonable underlying principles governing the robot's actions, be they appropriate or misplaced.

In contrast, experience shows that even the most simple combination of an underlying desire and affective state modulating the otherwise straightforward functional behavior of the robot significantly changes the way people interact with the robot, attributing a much higher degree of "understanding" and "realism" to the inanimate object.

Robots will clearly not have as rich an intentional mechanism as humans do, but there is value in applying a objective-based system in order to achieve more complex behavior patterns. If we are attempting to create interactive robotic actors, the notion of objectives and desires must exist at the base of the robot's behavior, even if they are based merely on a very rudimentary motivational system.

More specifically, the idea of an "inner monologue" is a powerful one that can, and should, be transferred to the realm of Human-Robot Interaction. Just as an actor should be constantly conducting an inner monologue to achieve continuity and realism in his behavior, and to make his externally evident actions emotionally based, it makes sense for an HRI-centered robot to continuously progress internal processes, instead of "waiting" for the human and then selecting the correct response.

Endowing an HRI robot with "inner monologue" might have a similar effect on robots as it has on actors, i.e. result in a much more natural and continuous interaction, as the robot picks up impulses (in the Meisner meaning of that word [Meisner and Longwell, 1987]), issued by the human before it actually responds to those impulses. This could be a significant key in avoiding the currently prevalent command-and-response behavior robots usually display, and lead to more fluency in human-robot joint action.

With inner monologue, sparse actions can become tips of icebergs of internal processing, rather than a disconnected array of floating islands of behavior.

Inner monologue must, of course, play out externally—cueing the human counterpart as it evolves, leading up to and filling up blank spaces between actions, and adding a layer of nonverbal communication to human-robot interaction.

8.5 Summary

Actors face some challenges that are surprisingly related to the ones HRI designer are struggling with. This paper identifies three persistent threads appearing throughout most acting theory and practice literature. While not providing resolute answers, they point us in interesting directions. Moreover, their pervasiveness may indicate that they hold central ideas that make acting possible, rendering them candidates for serious consideration by HRI designers.

First, we reviewed the notion of the *psycho-physical unity*. In today's interactive robots, behavior systems and motor systems are separate, if communicating, units. Acting method suggests to unify the two to be aspects of the same process. Second, *mutual responsiveness* has been shown to be a cornerstone of—among others—Meisnerian method. Focusing on the *between* instead of the *within* could prove to be a promising direction for rethinking HRI design. Third, Stanislavkian teaching suggests analyzing action in terms of *objectives and inner monologue*. Taking the 'tip of the iceberg' approach to Human-Robot Interaction might make human-centric robots behave in a more principled, fluent, and continuous fashion.

On a more practical note, actors' preparatory exercises aimed at achieving these goals could also hold valuable clues for our robotic endeavor.

Chapter 9

Designing a Robotic Desk Lamp

Robot design is currently a predominantly grassroots activity, performed by a loosely selected collection of graduate students, mechanical, electrical, and computer engineers, with the occasional input from professional designers in related fields.

We believe that robot design for HRI is bound to express itself as a separate field charting an interdisciplinary course on the brink of mechanical, electronic, product, human-factors, interaction, and animation design.

In this chapter we attempt to formally describe the design process of the robot AUR, a robotic desk lamp, which served as the platform for the implementation of the cognitive architecture laid out in the previous chapters.

9.1 Motivation

The motivation to design a robotic desk lamp is threefold:

First, it stems from a commitment to explore the expressive and interactive aspects of simply moving, non-humanoid robots (see also: Chapter 10). Non-humanoid robots allow us to consider our relationship to mechanistic artifacts, allow for a more exploratory design, and define a feasible entry point of robots that might be implemented on a mass-market scale in the near future.

Then, a robotic lighting assistant, collaboratively illuminating a scene for a human working on a bi-manual construction task has promising practical uses. In human collaboration it is often the case that an assistant to a person engaged in a task aids the performance by illuminating the task-relevant area. The prospect of a robotic illuminating assistant lighting the right thing at the right time solves a recurring real-world problem, which — at the moment — does not have a satisfactory automated solution. Advances towards this goal could be of interest to a number of real-world applications, such as space station maintenance, deep sea exploration, surgery, machine repair, and others. Also, much like in the case of a human lighting assistant, the robot's light beam can be used as a deictic gesturing device, pointing out aspects of the task that the operator needs to be aware of.

Finally, a collaborative lamp serves our research agenda well. To appropriately direct a lighting beam towards a shared task location is a highly embodied mission which takes into account the actions and goals of the human team member, their perspective on the task, and provides an interesting space for improvement by practice, as well as the employment of anticipatory actions.

An early prototype of this concept included one actuated degree of freedom (DoF) at the neck, one passive DoF in the head, a variable-intensity white light, a proximity sensor at the base, and a capacitive sensor on the head (Figure 9-1). While the robot's behavior was no more than a simple three-state machine, reacting to the lamp's immediate environment, the expressive qualities derived from this simple embodied artifact were promising.

9.2 Degrees of Freedom

The current lamp design extends the original prototype in several ways: instead of a one-DoF body, the lamp is utilizing a five-DoF robotic arm, adapted from a previous robot, RoCo, a robotic computer [Breazeal et al., 2007]. Two DoFs are at the base, moving the

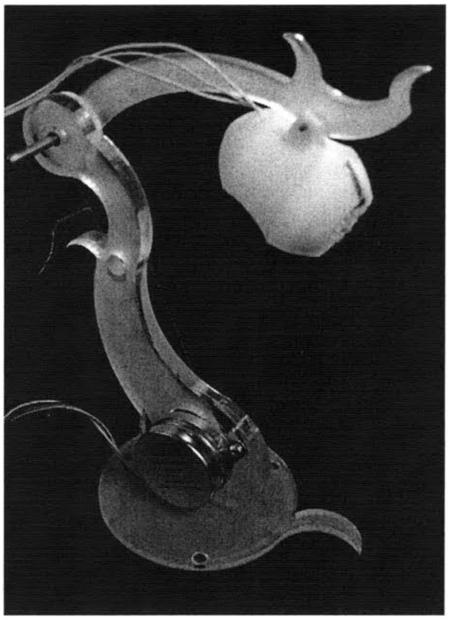


Figure 9-1: Early prototype of a robotic desk lamp.

robot's neck (base yaw and base pitch) and three DoFs are at the top of the neck, moving the head attachment (head yaw, head pitch, and head roll) . An MEI motion controller is used to drive five harmonic drive geared DC motors with optical encoders. Figure 9-2 shows the robotic arm used for the desk lamp.

The lamp's lighting element was extended from a single color intensity-varying light to a Red-Green-Blue controlled LED lamp. In addition, the lamp's light cone diameter was designed to be able to change both as a functional and a communicative element. We have chosen to include non-human modes of expression — changing color and light beam — in addition to anthropomorphic posture, to be able to evaluate the effects of non-humanoid nonverbal behavior in human-robot interaction.

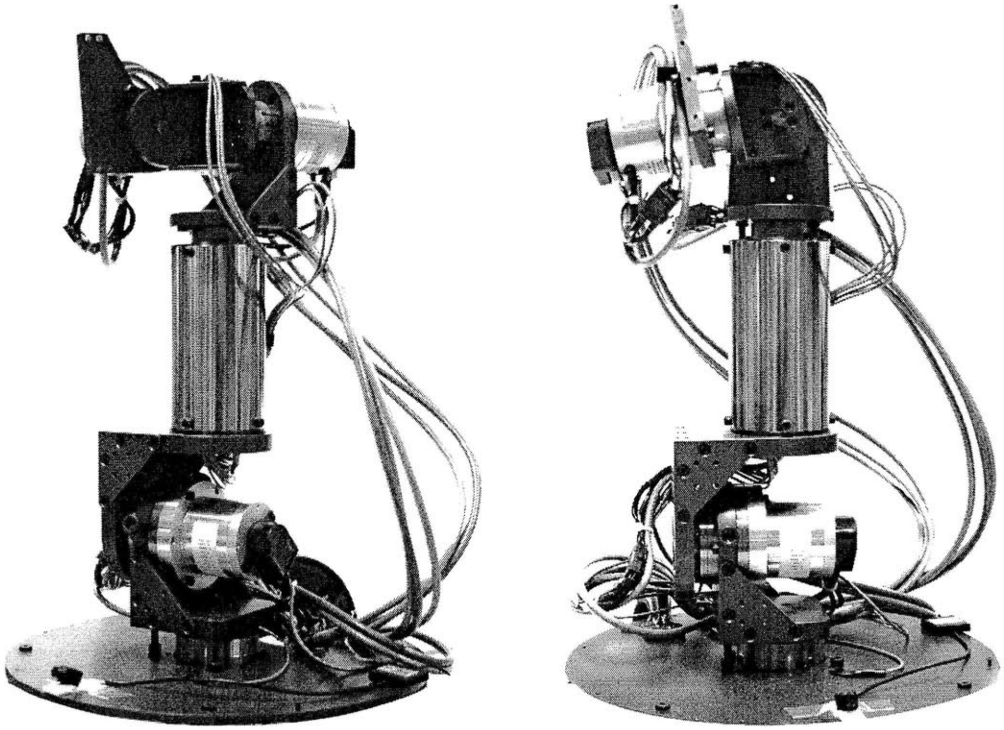


Figure 9-2: 5 degrees-of-freedom robotic arm adapted from the RoCo robot.

9.3 Two-Part Evolutionary Prototyping

The design process of the lamp follows a path of two-part evolutionary prototyping (Figure 9-3). Each stage in the form design led to the following mechanical design, in turn informing the next form design stage.

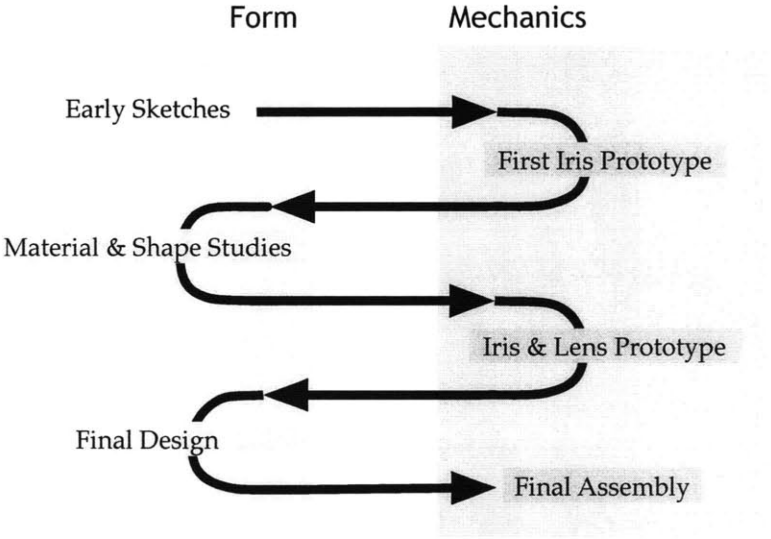


Figure 9-3: Two-part evolutionary prototyping approach taken in the lamp design process.

Early sketches were based off of the existing robot body, exploring possible morphologies. Figure 9-4 shows four examples of these studies.

To allow for a varying size light beam, we then developed the first-generation robotic aperture to control the beam size. The aperture consisted of 8 blades locked between two counter-rotating plates and placed in front of the lighting element (Figure 9-5).

In this initial prototype, the front plate was rotated using an external gear drive with potentiometer position control, and held in place by two idle rollers. Figure 9-6 shows the fully assembled first prototype. The blades in this prototype were cut out of paper, while alternative materials were explored, such as teflon, aluminum, acrylic, and stainless steel — which was the eventual material of choice.

The inclusion of the aperture mechanism, and the choice of drive and feedback compo-

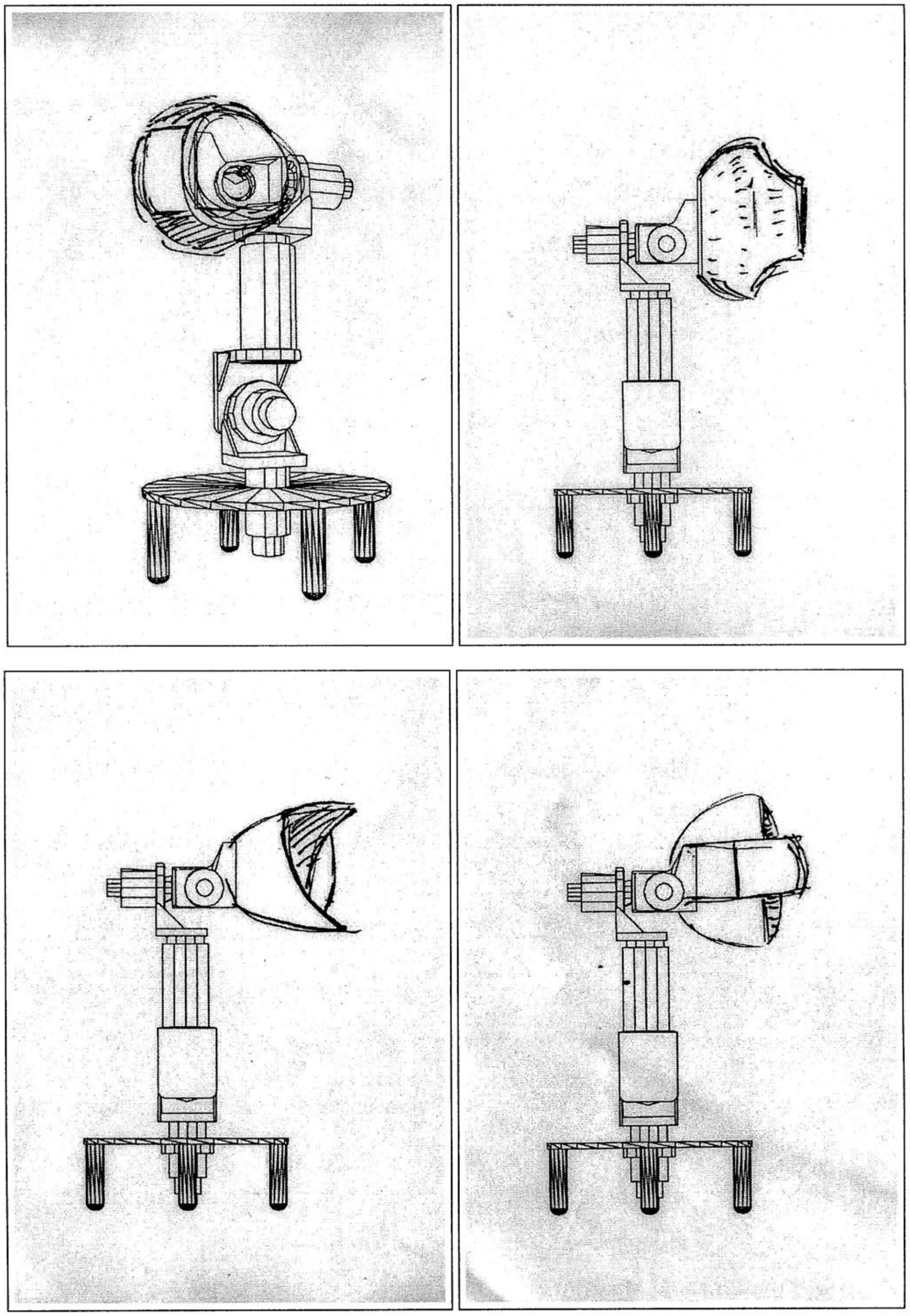


Figure 9-4: Early sketches of robotic lamp morphology studies.

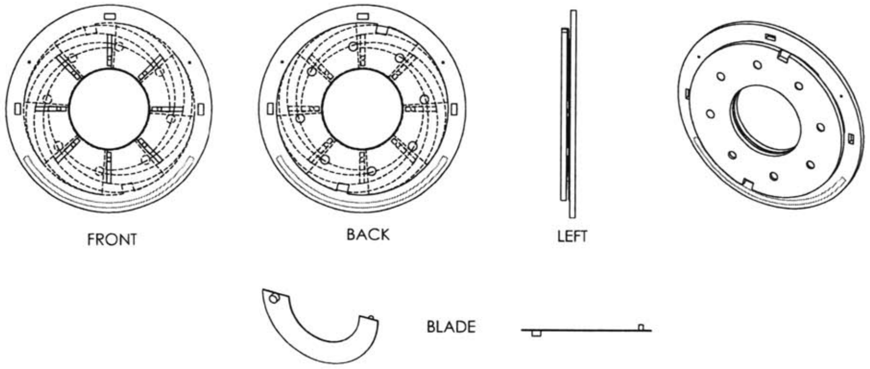
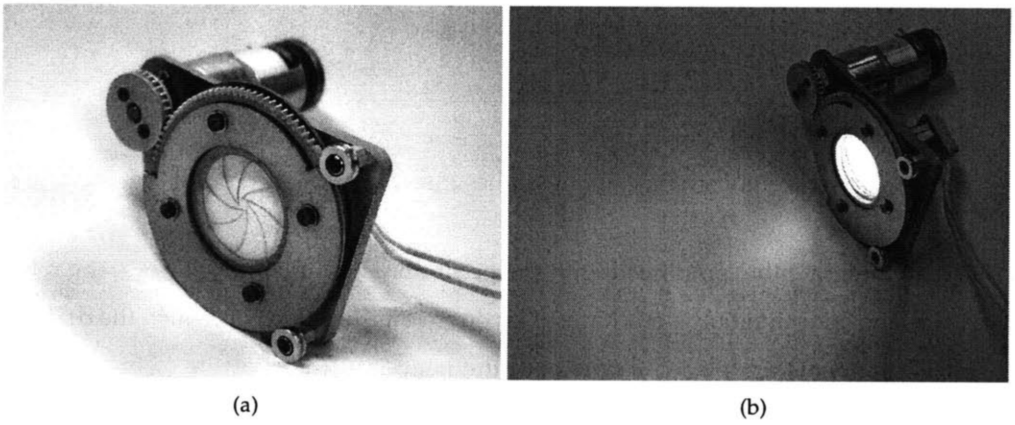


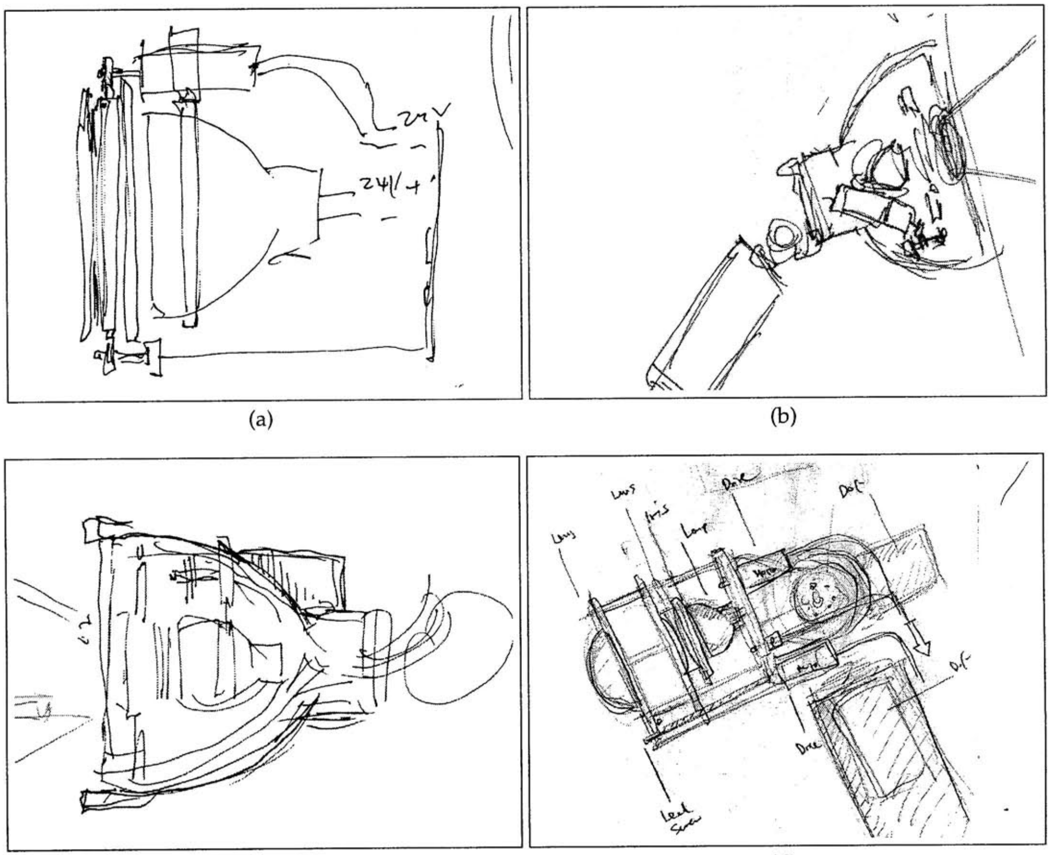
Figure 9-5: Iris made up of eight blades locked between two counter-rotating plates.


(a)

(b)

Figure 9-6: First mechanical prototype of the robotic iris; (a) closed, and (b) open.

nents informed the next iteration of design, seen in Figure 9-7. Some studies include a lens, positioned to define the edges of the lamp's beam.



(a)

(b)

(c)

(d)

Figure 9-7: Drive mechanism placement studies informing second round of form design.

A desire to integrate the lamp head with the predominantly tubular design of the robotic arm, led to the adoption of the mechanism layout in Figure 9-7 (d), in which the drive and sensing components are placed *behind* both the lamp and the aperture. This was made possible by rotating the lamp together with the back plate of the aperture and driving the assembly through an internal gear.

Figure 9-8 shows a prototype of this new drive mechanism, with motor, sensor, lamp, aperture, and lens in place.

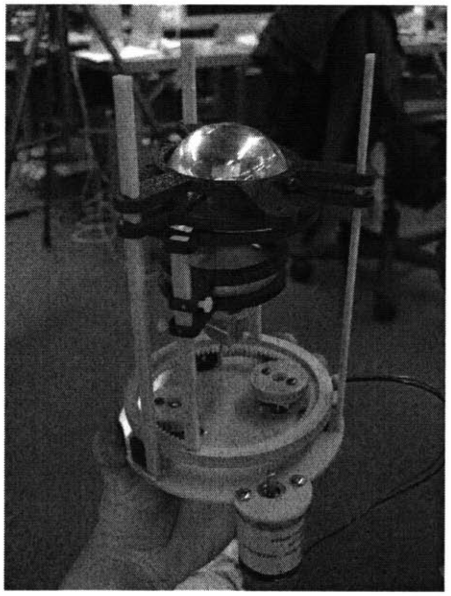


Figure 9-8: Second iteration drive mechanism.

This implementation of the lamp's mechanical design led to the consideration of a number of creature-like designs (e.g. in Figure 9-9), which were rejected for a form resembling a traditional lampshade. This decision is in line with the argument set forth in Chapter 10.

Settling on the overall lamp design — a roughly lampshade formed head with a color-controlled light, actuated aperture, and lens in a linear layout — materials were considered next. To enhance the readability of the lamp's color modality from multiple directions, a diffuse material, such as sandblasted acrylic or vellum paper was considered.

Since the lamp head was to be mounted on an existing arm, made of part blue-anodized, part nickel-plated aluminum, a nickel-plated element in the lamp was thought to be able to bridge the material gap between the white semi-transparent end effector, and the body makeup. Figure 9-10 exemplifies these material and light considerations in of a series of morphological studies.

Once the materials, shape, and components were set, we embarked on a series of three-dimensional morphological studies constrained by these selections, and leading to the final lamp design. Figures 9-11 through 9-14 show a selection from these 3d model sketches.

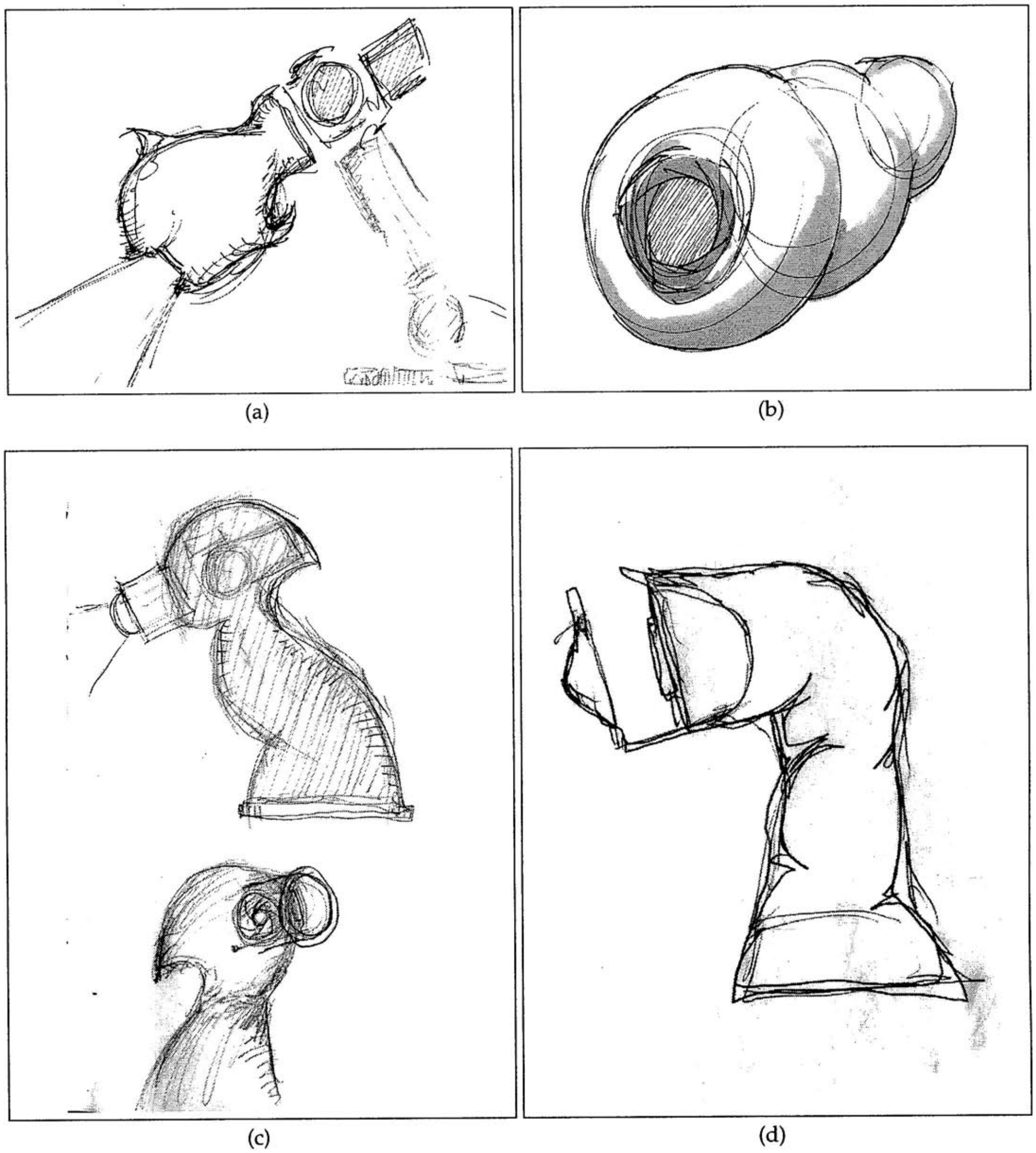


Figure 9-9: Intermediate creature-like designs.

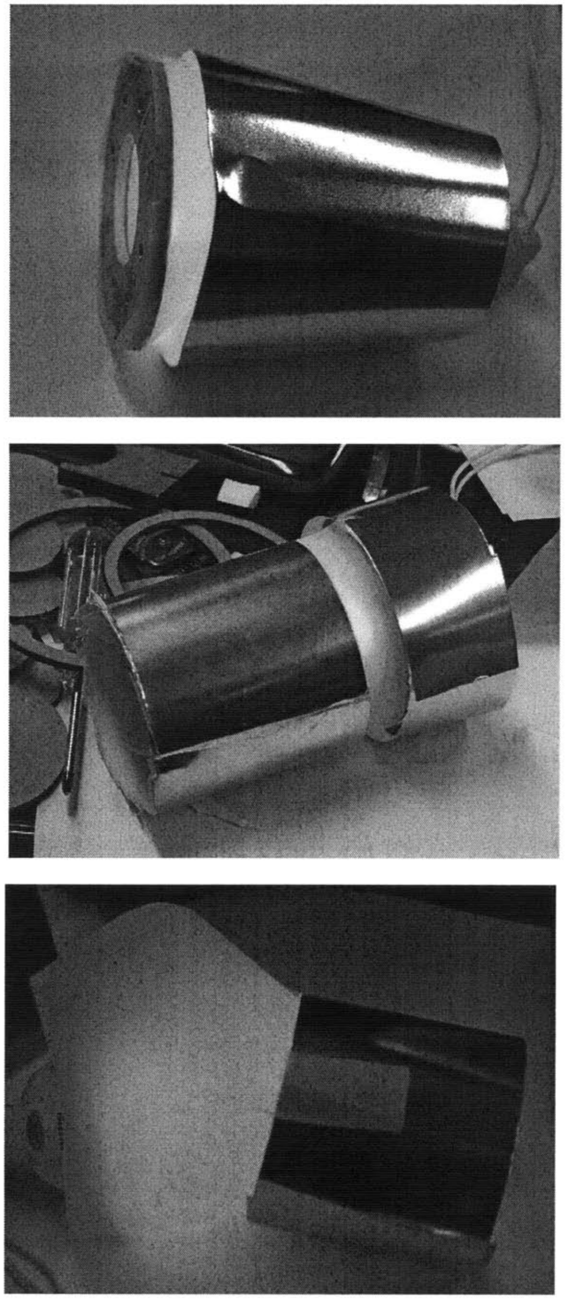


Figure 9-10: Material and light studies.

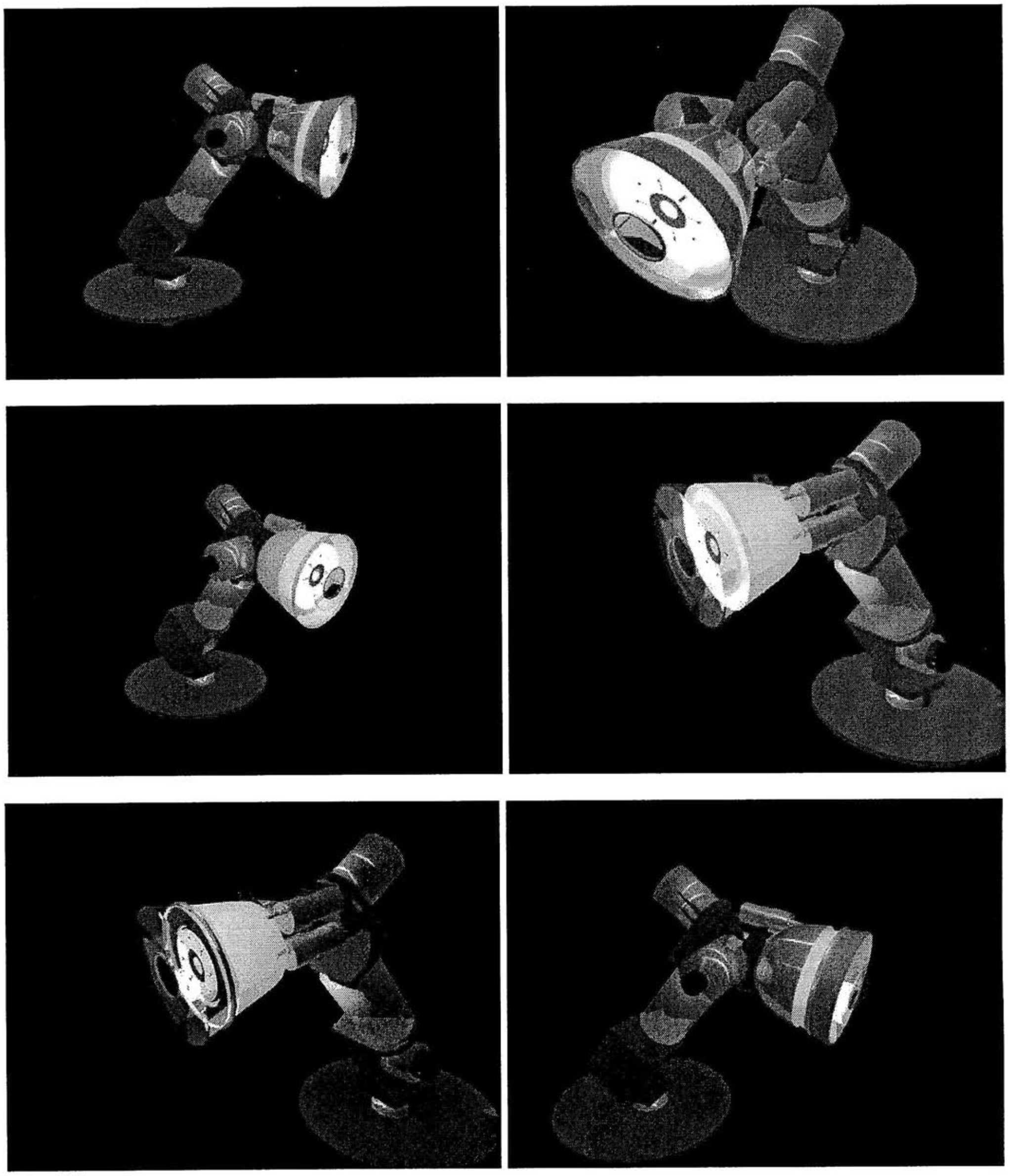


Figure 9-11: 3D models of final design variations.

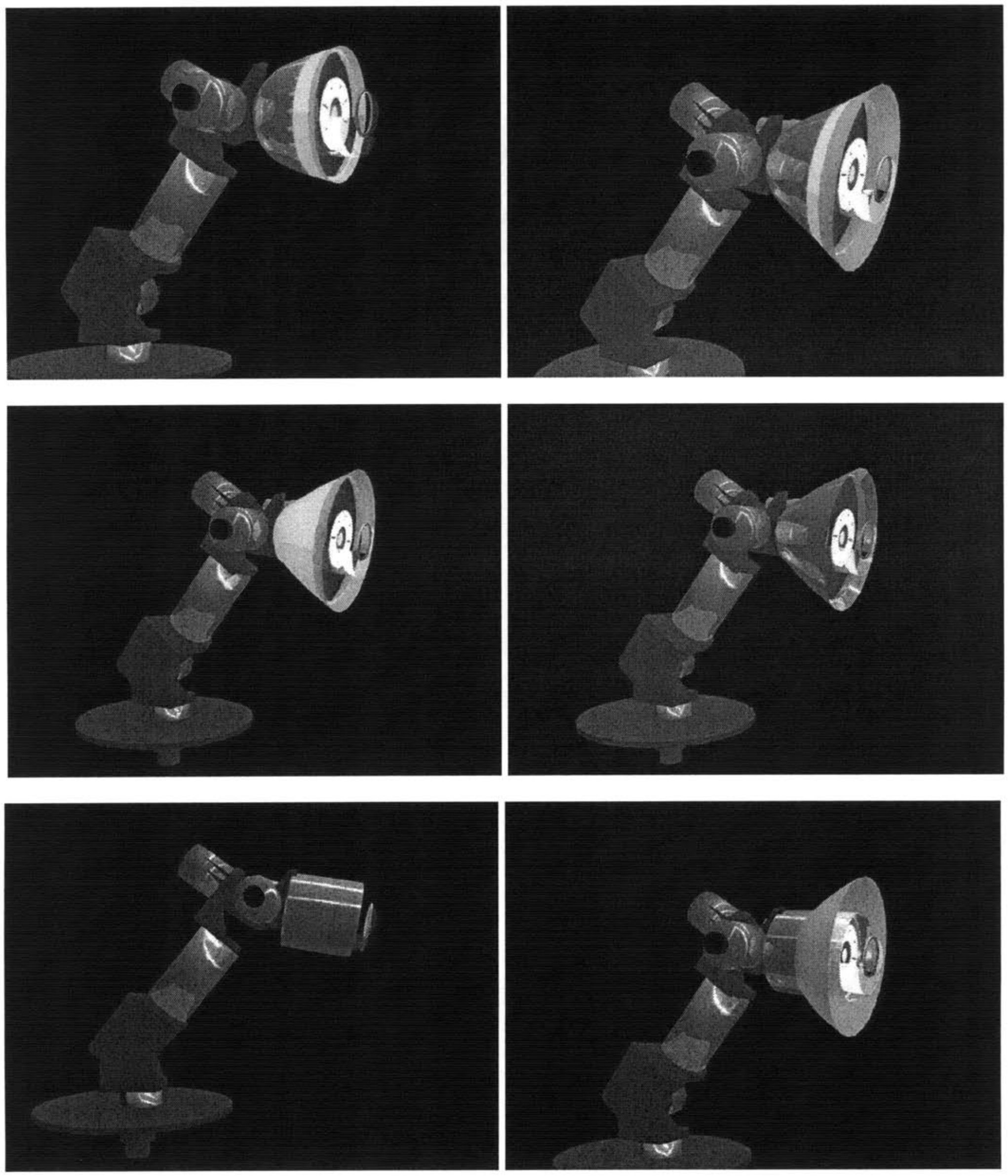


Figure 9-12: 3D models of final design variations (cont.)

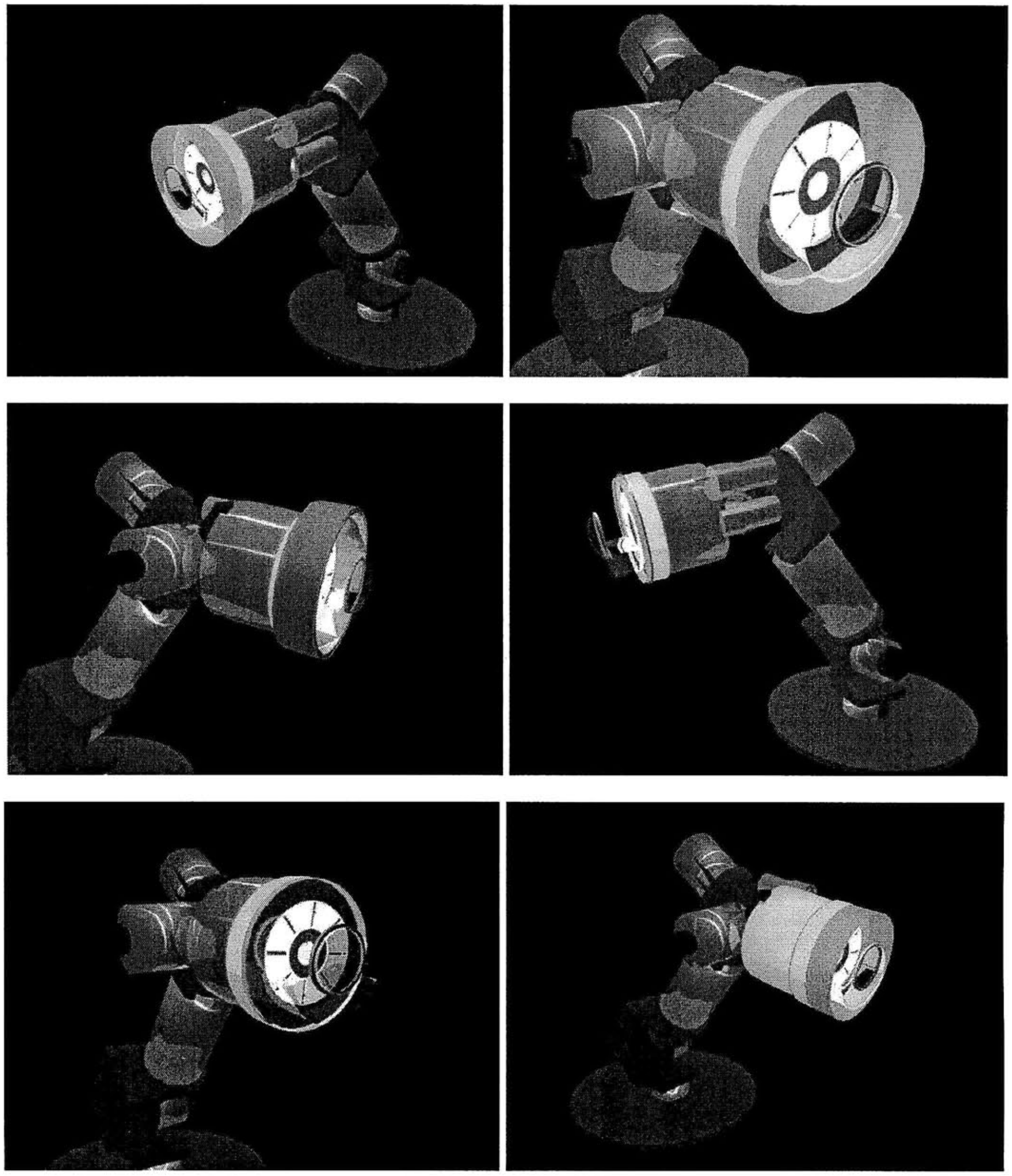


Figure 9-13: 3D models of final design variations (cont.)

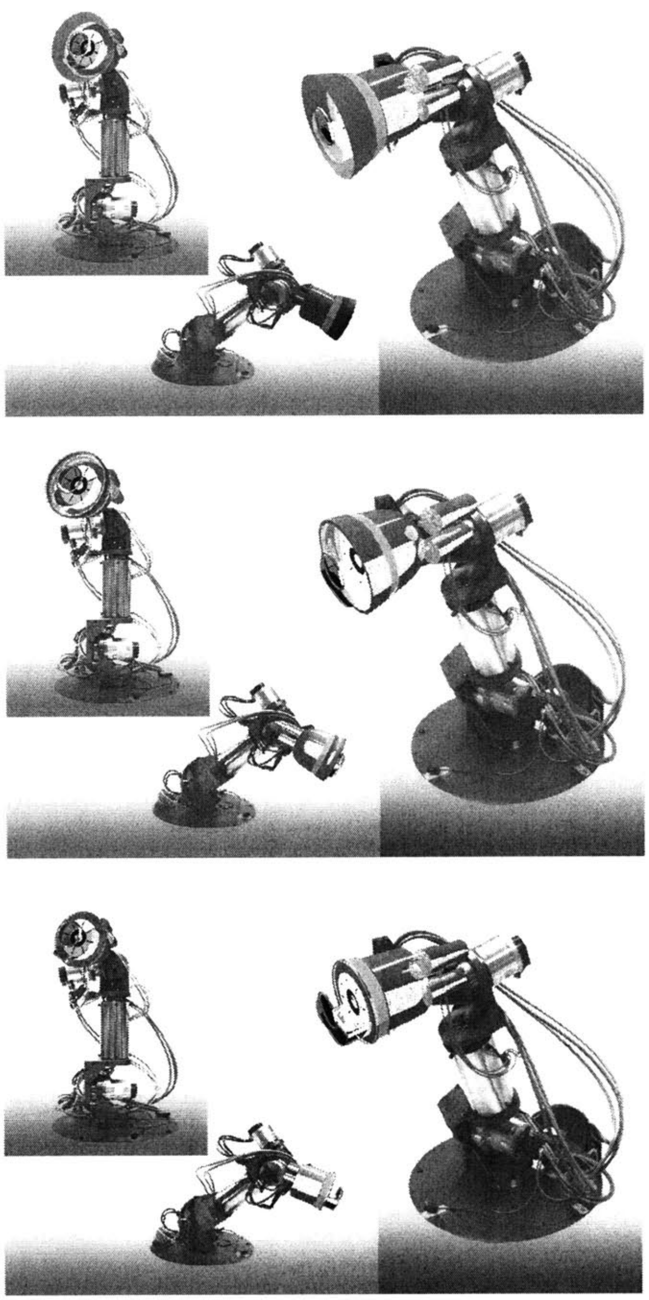


Figure 9-14: 3D models of final design variations on actual robotic arm.

9.4 Final Lamp Mechanical Design

The mechanical design of the final lamp prototype consists of three major components:

- **Drive train**, including the interface to the robotic arm,
- **Lamp rotor**, with the aperture back plate
- **Shell**, with the aperture front plate

9.4.1 Drive train

The drive train includes the interface to the robotic arm, and the lamp's base plate, holding the DC motor and the potentiometer used for position feedback control. Both components are attached to 48-pitch 1″ gears. The base also contains a centered ball bearing (Figure 9-15).

Figure 9-16 shows an axonometric diagram and a photograph of the drive train. Note that in the photograph, the rotor shaft is shown inserted into the bearing.

9.4.2 Rotor

The rotor (Figure 9-17) is made up of a custom-machined drive shaft held in place at the bearing with a shaft collar. A base plate attached to an internal 48-pitch 2.75″ diameter gear is moving two drive rods. The lamp element is attached to the drive rods, as is the back plate of the aperture, enabling the blade disposition resulting in changing the aperture size.

Figure 9-18 shows rotor base components (a) , as well as the assembled rotor base, out-of (c) and in-place (b) on the drive train. Figure 9-18 (d) shows the fully assembled rotor in place.

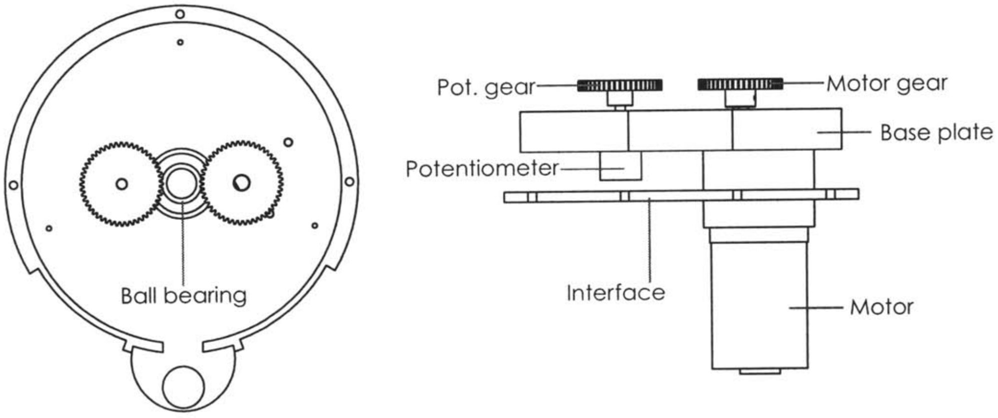


Figure 9-15: Diagram of the lamp's drive train.

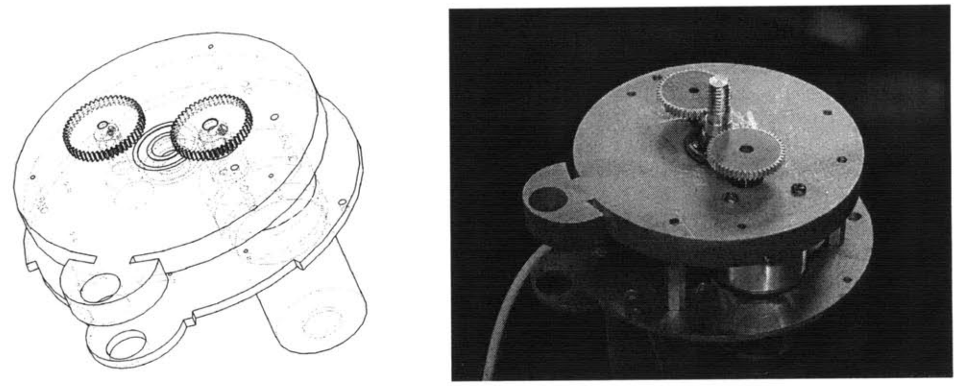


Figure 9-16: Diagram and photograph of the lamp's drive train.

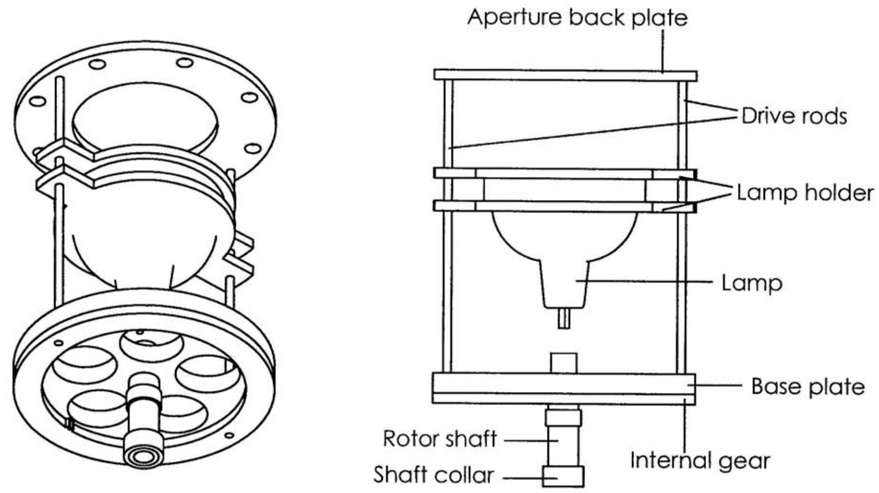
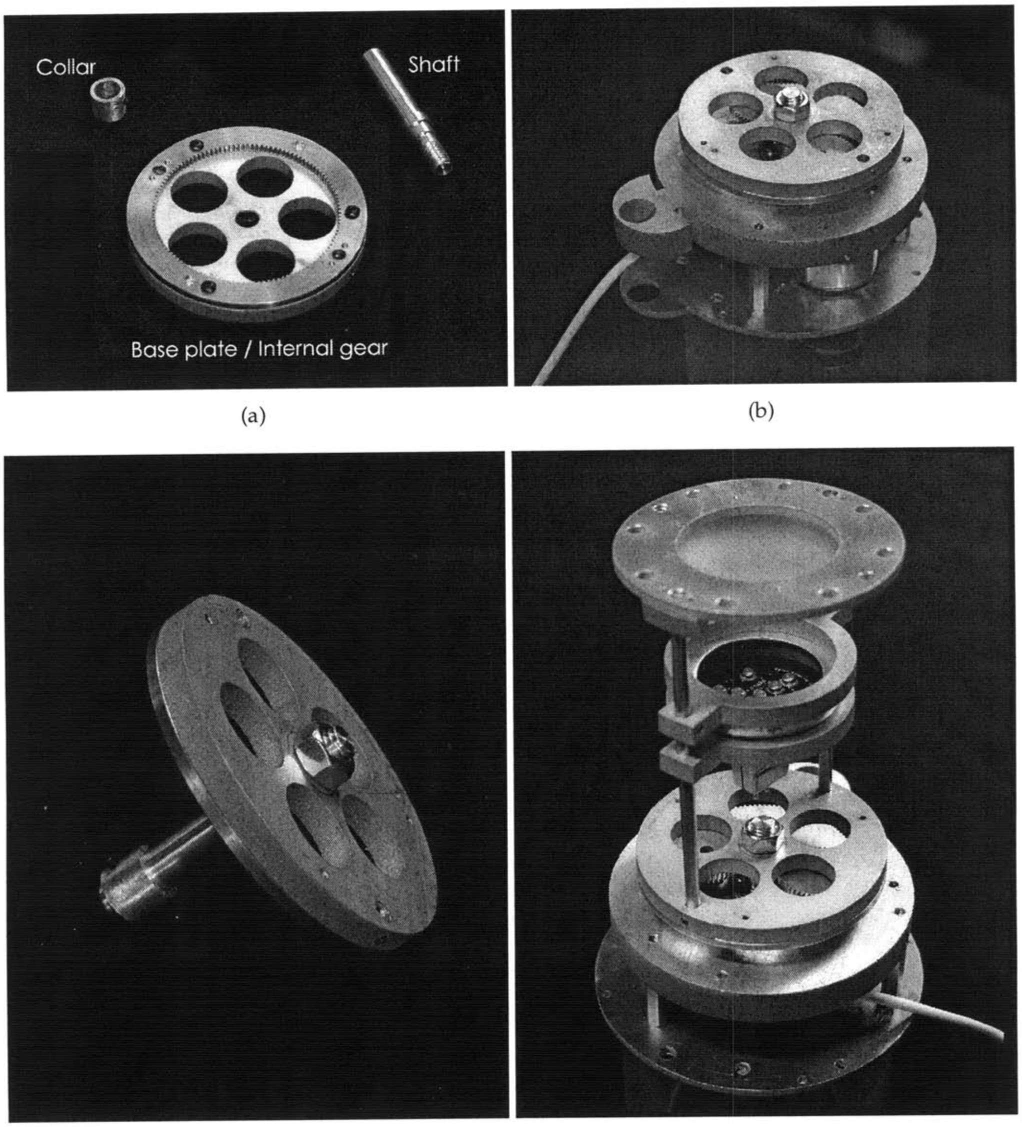


Figure 9-17: Diagram of the lamp's rotor.



(a)

(b)

(c)

(d)

Figure 9-18: Photographs of the lamp's rotor.

9.4.3 Shell

The shell performs two functions: aesthetically it diffuses the lamp's color through the plastic shade enabling this modality to be viewed from more than just the frontal point of view. The nickel plated aluminum tube functions as a formal bridge between the robot's body and its role as a lighting device.

Mechanically, the shell holds the aperture's front plate in place, entrapping the blades between itself and the rotating back plate, resulting in the opening mechanism.

Incidentally, we also found that the lamp's shell functions as a shock absorbing element in case of a mechanical failure of the lamp arm.

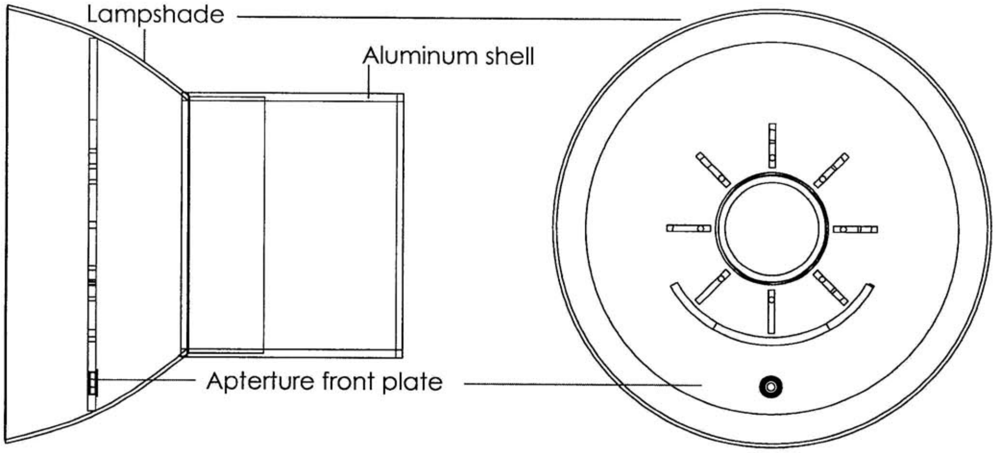


Figure 9-19: Diagram of the lamp's shell.

Figure 9-20 shows the full lamp assembly in axonometric projection. Figures 9-21 through 9-24 show photographs of the assembled lamp.

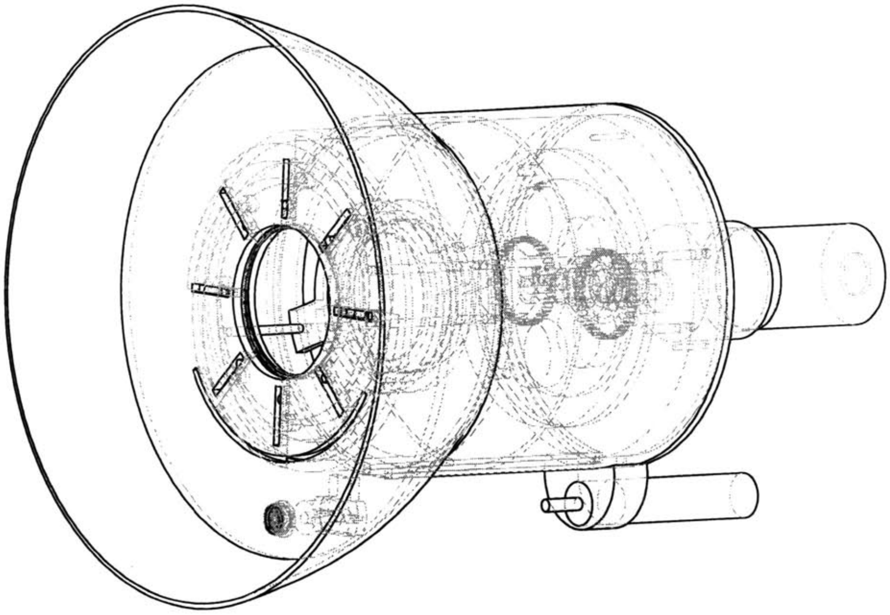
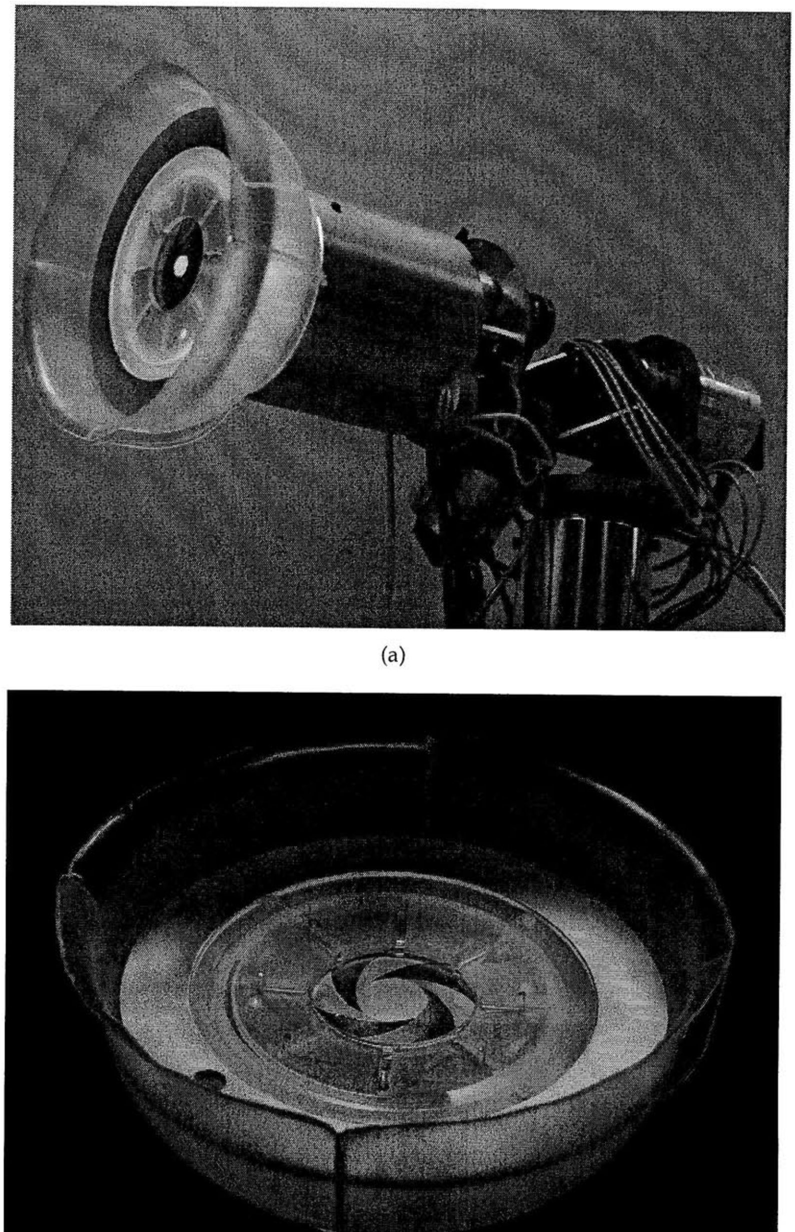


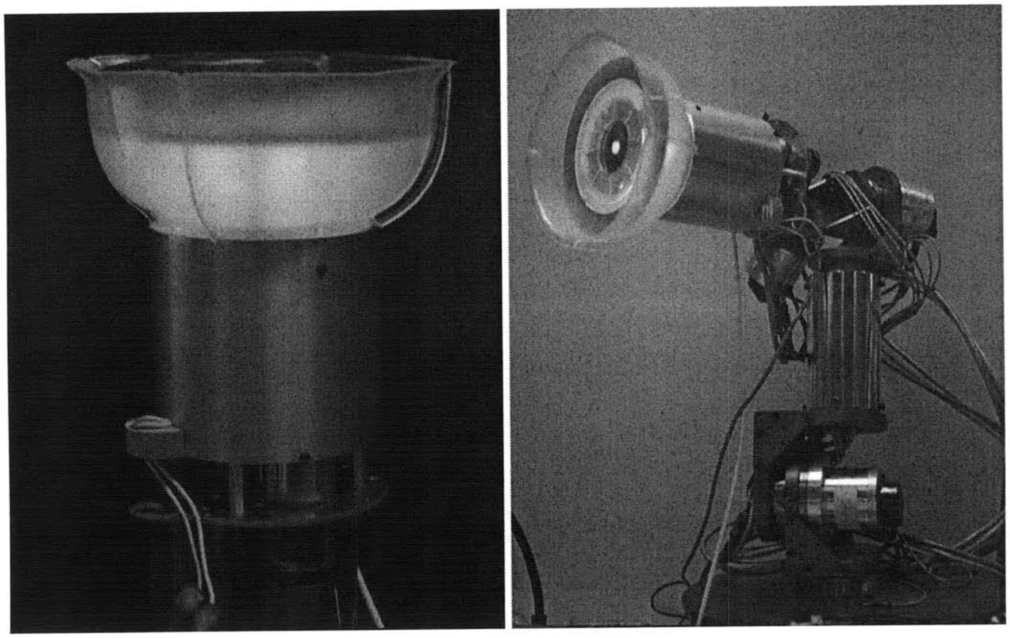
Figure 9-20: Diagram of the full lamp assembly.



(a)

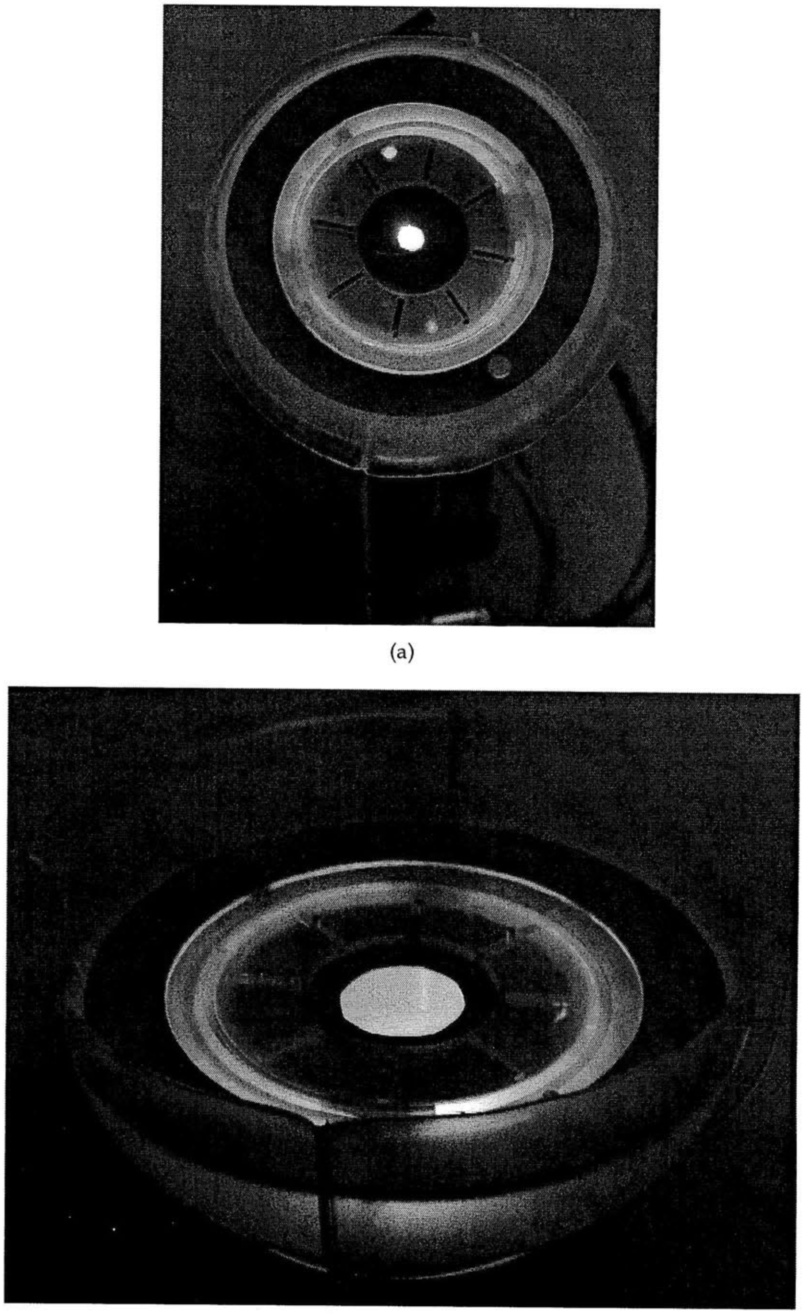
(b)

Figure 9-21: Photographs of the final assembled lamp.



(a) (b)

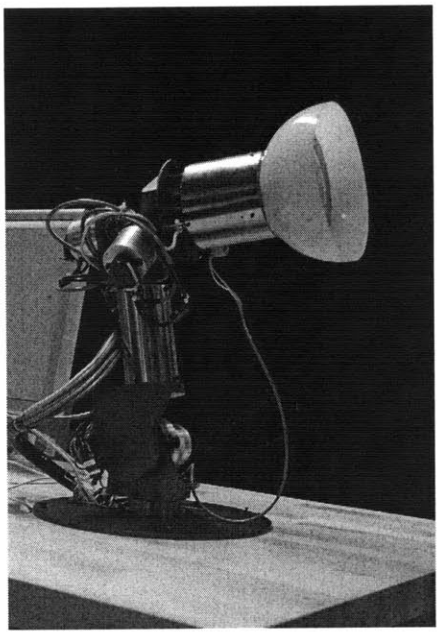
Figure 9-22: Photographs of the final assembled lamp (cont.)



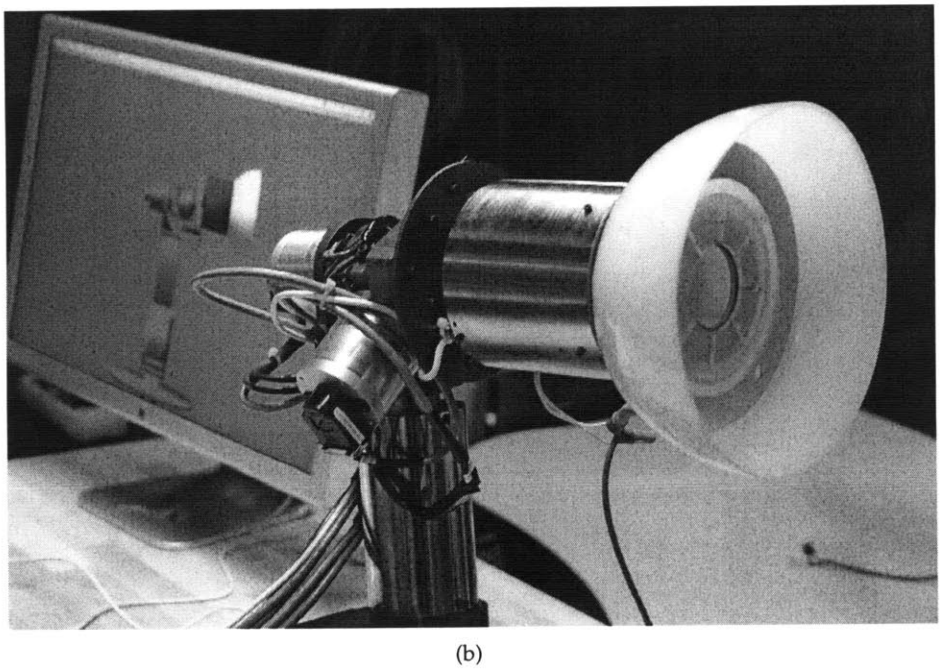
(a)

(b)

Figure 9-23: Photographs of the final assembled lamp, showing the closed and open aperture.



(a)



(b)

Figure 9-24: Final assembled lamp with alternative lampshade and simulator in background.

Chapter 10

On Non-Humanoid Expressive Robots

The platform serving to illustrate the core implementation of this work's research (described in Chapters 6 and 7) is a non-humanoid robot, a robotic desk lamp. Its design process is described in Chapter 9. In the following chapter we will briefly outline some of the considerations for pursuing a path of non-humanoid design for expressive robots in human-robot interaction.

10.1 Non-antrhopomorphic character design in art

Be it for reasons of technological constraints or artistic experimentalism, artists have often attempted to create valuable works within frameworks of imposed limitations. At times these constraints had a surprisingly positive impact on the final product, proving to be not hurdles but rather inspirations.

In the field of character animation, this principle is classically associated with "The Dot and the Line" [Jones, 1965], in which complex narrative and character development are implemented using two very simple characters, a dot and a line (Figure 10-1). All expression is performed in motion, timing and staging alone. While technically impressive, this film is also often quoted for distilling the core principles of expressive cartoon "acting" without other visual aids.

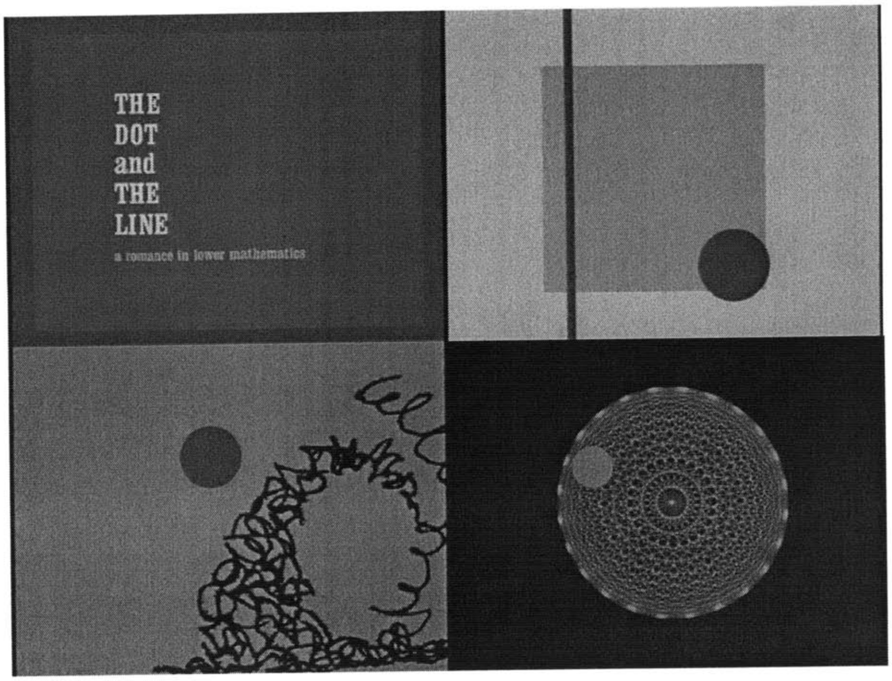


Figure 10-1: Screen capture images from Chuck Jones' 1965 short "The Dot and the Line".

More recently, in a similar example of utilizing technical limitation, Lasseter [1986] produced "Luxo Jr." which features two simple desk lamp characters. For lack of appropriate 3D modeling and animation software, Lasseter used only basic 3-joint rigid characters and, once again, all character expression and narrative is implemented trough motion in these 6 joints, with heavy reliance on sound cues. This short continued to become a textbook example of narrative 3d animation using simple motion and sound effects. "Luxo Jr." also served as one inspiration for the design of the robotic desk lamp in this dissertation.

10.2 Non-humanoid expressive robots

It is similarly worthwhile explore robotic characters digressing from humanoid and anthropomorphic modeling, and instead focus the design effort on simplicity and motion. We also urge robot designers to consider familiar objects and embellish them with actuation and animation. Several reasons for this have been discussed in the past [Duffy, 2003], including the exceeding expectations one has of a robot bearing a human form, the need for emphasis on behavior over appearance, and the avoidance of the so-called "uncanny valley".

We also believe that the evaluation of non-anthropomorphic robots, and in particular familiar objects which were embellished with robotic traits, encourage people working with such robots to consider their relations to such "moving inanimate" objects, as opposed to building upon the preconceptions brought to the interaction by previous human-human interaction.

We would like to offer two additional design considerations:

10.2.1 Freedom of exploration

A traditional design path in robotics imitates the human physique and features, building robots that usually have one or more of: arms, eyes, mouths, a head, or a human-like torso. Much of mobility research is also concerned with the imitation of human-like legged locomotion (e.g. Sakagami et al. [2002]).

This approach, based on the implicit assumption that the human form and our behavior are a pinnacle of design or evolution, inherently falls short. It views the human example as a practically unattainable goal towards which small advances can be made, bringing the robotic simulacrum a little closer to the human ideal. Starting with a void, the direction of such imitation is clear and constrained by the human example. Its merit is strictly evaluated with respect to the human original.

In contrast, starting the design process with an existing artifact allows a certain freedom to explore different avenues of embellishment. Without a set ideal guiding the robot design, the same familiar object can be imagined through a variety of evolutions. In our example, a robotic desk lamp superseded a regular lamp by allowing the light beam to change color and using an iris to focus the beam. We defined five degrees of freedom and their extent, thus defining the movement capabilities of the lamp. By specifying the trajectory of the lamp's movement, we design the robot's attitude and its implied "personality", in the character animation meaning of the phrase. All of these are not constrained by a "robotic lamp ideal", but instead grow out of the familiarity of the robot's ancestor, the household desk lamp. This lack of preconceived ideal allows for more creative freedom in the robot design process.

10.2.2 Market feasibility

The second driver to consider HRI research on non-anthropomorphic expressive robots is that of market feasibility. While humanoid robots are developed in many research laboratories, their advent into the consumer market is still unforeseeable.

Considering simpler non-humanoid forms with fewer degrees of freedom holds the promise to study human-robot interaction that is relevant to more imminent consumer products. To date of this writing, 5-DoF robots, including their mechanical design and control software, are not far-fetched possibilities for mass-manufactured products.

In addition, focusing robot design around familiar existing household objects may allow for more rapid market acceptance, as these artifacts' utility in people's homes is well-established, while the introduction of humanoid robots is bound to require a paradigm shift before becoming common practice.

Additional merits from HRI research in simple, non-humanoid expressive robots include the possibility of exploring human reaction to basic physical animation, with the opportunity to systematically explore simple controlled affective response. Moreover, such robots boast a cost-effective design, allowing for evolutionary prototyping, frequent re-implementation, and large-scale replication.

Part IV

Closure

Chapter 11

Contributions

We see our main contributions to the field in four realms: *cognitive modeling and implementation; empirical study and metrics; design;* and *theory*. This chapter summarizes these contributions.

11.1 Cognitive modeling and implementation

This work developed a perceptual symbol based computational cognitive model including several key features:

Perception-action integration True to the belief that perception, decision, and action selection must exist in a connected, mutually influential relationship, our cognitive model views perceptual processing and action as similar and interconnected entities in a continuously dynamic activation network.

Modality streams Modeled after Damasio's *convergence zones* [Damasio, 1989], our perceptual processing is organized according to modality-specific streams, but allowing for intermodal connection.

Simulation and top-down processing Perceptual processing occurs not only from the sensory to the concept and action-selection layers, but can be triggered from any point in the cognitive network, enabling sensory-independent top-down bias of perceptual streams.

Anticipation and emulation Bayesian learning is used to model practice through the triggering anticipatory top-down simulation, akin to Wilson and Knoblich's *emulators* [Wilson and Knoblich, 2005].

Co-occurrance reinforcement Dynamic feedback reïnforcement between processing regions models one kind of long-term memory in consistently co-occurring perceptual activation patterns.

We have implemented this cognitive model in software and developed a visualizer to track activation in this network. We showed the application of our system on two different robotic platforms, a humanoid robot and a non-anthropomorphic robotic desk lamp, and for two different human-robot collaborative tasks. We have implemented the anticipatory practice element of this architecture on a third task, run on a simulated robot.

11.2 Empirical study and metrics

Two human-subject studies were run to corroborate our approach. One, including 32 subjects, evaluated the effects of only the anticipatory practice portion of our model. A second study, including 33 subjects, compared the use of the emulator approach to a pure bottom-up version of our cognitive model. We showed significant differences on a number of behavioral metrics and self-report evaluations.

As part of these empirical studies, we developed a number of behavioral fluency metrics that could well be related to the notion of fluency in human-robot interaction. These are

Concurrent motion The percentage of time in which both human and robot were in motion;

Human idle time The percentage of time in which the human was not active, or not in motion;

Robot functional delay The time lapsed between the human's action and the robot's subsequent action.

We found two of the three to be different between the two experimental conditions in the first study, and two to be different in the second study.

We have also developed a number of survey instruments appropriate for the study of human-robot fluency, and human-robot teamwork in general, and evaluated their reliability. These include:

HRT-ENJOY — The overall enjoyment of the teamwork experience;

HRT-FLUENCY — The sense of fluency in the teamwork;

HRT-IMPROVE — The sense of team improvement over time;

HRT-ROBOT-CONTRIB — The rating of the robot's contribution to the team;

HRT-ROBOT-TRUST — The human's trust in the robot;

HRT-ROBOT-CHAR — The robot's perceived positive character traits;

HRT-WAI-BOND — The human-robot teamwork (HRT) Working Alliance bond subscale;

HRT-WAI-GOAL — The HRT Working Alliance goal subscale

HRT-WAI — The overall HRT Working Alliance.

11.3 Design

As part of this work, we proposed to consider robot design as a discipline in its own right, taking inspiration from mechanical engineering, product design, animation, and human interface design.

Following a two-part evolutionary prototyping process, we designed and built a robotic desk lamp on the basis of an existing robotic arm. The lamp featured a varying width beam spanning the RGB spectrum.

11.4 Theory

Finally, we believe to have made two theoretical contributions. First, we raised the subject of fluency in human-robot interaction, and encourage researchers in the field to consider rhythm, timing, and the quality of action meshing as an integral part of their evaluation of interactive robot intelligence.

We hope to have raised an interest in this unexplored aspect of HRI, and believe that the principles we have proposed for achieving this sense of fluency, and which we denoted "embodiment" — specifically the tight coupling between perception and action, the use of anticipation, and the principles of perceptual symbols and simulation — can hold keys to other aspects of robot intelligence as well.

Second, we proposed the notion that theater acting holds valuable insights into the analysis and generation of synthetic behavior, making it an appealing candidate to inform human-robot interaction, which — we argue — faces many of the same challenges as actors do.

We pointed out that acting theory underwent a metamorphosis from a symbolic to an embodied approach, not unlike the one we propose for HRI in this dissertation. As a result we identified three lessons from acting method: *psycho-physical unity*, *mutual responsiveness*, and *inner monologue*.

Chapter 12

Future Work

This dissertation makes no claim to completeness. Much of the initial steps indicated herein suggest more open questions and avenues for future work than the answers they provide. We attempt to point in some of those directions for extension in this chapter.

12.1 Intentions, Motivations, and Supervision

Much of the core architecture described in this thesis, as well as the implementations thereof, are predominantly apt for governing automatic or routine activity. Perceptions precipitate concepts, which in turn prompt actions leading to future concept, perception, and motor activation. While this model can be adequate for well-established behavior, a robot acting jointly with a human must also behave in concordance with internal motivations and intentions, and higher-level supervision. This is particularly crucial because humans naturally assign internal states and intentions to animate and even inanimate objects [Baldwin and Baird, 2001, Dennett, 1987, Malle et al., 2001]. Robots acting with people must therefore behave according to internal drives as well as clearly communicate these drives to their human counterpart. This view is also supported in modern acting literature supporting "inner monologue" as a pillar of objective-driven acting (see e.g. Moore [1968]).

In a separate publication [Hoffman, 2005b] we propose a supervisory intention- and motivation-based control that could be used to affect the autonomic processing scheme outlined above. At the base of this supervisory system lie core drives, such as hunger, boredom, attention-seeking, and domain-specific drives, modeled as scalar fluents, which the agent seeks to maintain at an optimal level (similar to Breazeal [2002]). If any of those fluents falls above or below the defined range, they trigger a goal request. The number and identity of the agent's motivational drives are fixed. Goals are represented as a vector indexed to the various motivational drives, and encoding the effect of achieving a goal on each of the drives. This representation can be used by a goal arbitration module to decide what goal to pursue. Under the assumption that goals are mutually exclusive at any point, the relative deviance of the motivational drive from its optimal value in combination with the expected effect can be used to compute a utility function for selecting each goal. The most useful goal according to this function is thus selected.

Once a goal is selected, a planner is used to construct a plan that satisfies the goal. A plan can then activate or override the automatic action selection mechanisms in the action-perception activation network, activating simulators and guiding perception and motor activity, similar to the principled fashion laid out in Cooper and Shallice [2000]. The outcome of actions is fed back through perceptual acquisition to the motivational system.

Note that this model can also be used as a basis for an experienced-based intention reading framework, as outlined in the above-mentioned technical report.

12.2 Anticipatory Action Selection

As described in Chapter 3, several improvements to the anticipatory action selection should be considered.

Domain-specific knowledge can inform the action space at each decision point and thus bias the probability distribution of subsequent states. A discount factor in the learned state transition distribution, making more recent moves by the human teammate more salient to the robot than older experiences, can accommodate for learning time on the human's part, and make the system more robust.

The agent should be able to estimate of the human's turnaround time h, instead of using a fixed number. Ideally, this variable should be state-specific.

Additionally, the effects of anticipatory action vis-a-vis an expert — instead of a naïve — human teammate, is of interest, as is a controlled evaluation of the effects of the proposed fluency metrics on the efficiency of the task and the perceived fluency and commitment of the robot. A power analysis on the non-significant differences could indicate whether the found disparities between the two conditions can be statistically based with more subjects.

We would also like to measure the effects of practice in a human-human team solving the same task and compare the results from that study to the human-agent teamwork which we have studied in this work.

12.3 Perceptual Simulation Framework

Similarly, we would like to evaluate the same task that was used in the human-robot experiment on a human-human team, comparing the same metrics and thus gauging the verisimilitude of our practice mechanism.

Virtual "lesion" studies are appropriate, singling out the different mechanisms contributing to the robot's fluency, and evaluating each of these mechanisms' contribution separately. Also, *within*-subject studies of the concepts described here could shed a clearer light on some of the questions raised in the previous chapters.

We have yet to address the process that leads to the existence of so-called "privileged loops" between perception and action, and whether these should be learned, or pre-exist in the system. A clearer separation of the different roles of each of the process node types is also called for.

As mentioned above, the systems described herein are mainly concerned with short-term learning and practice. It makes sense to include the notion of long-term memory in the context of perceptual simulation, and explore the interrelation between such a long-term storage mechanism and the perceptual simulation framework advocated herein. In the past we have also proposed memory decay to be based on the amount of storage needed to hold a certain kind of memory. This notion has yet to be explored.

It would be interesting to implement the methods raised by Pentland [2004] in order to evaluate human-robot fluency, and to compare it with human-human fluency.

Finally, a number of open questions has been posed in Section 5.7. We will not repeat the issues raised in that section here, but do consider them as fertile ground for future research.

12.4 Robotic desk lamp design

Further evolution of our robotic desk lamp design is in place. We propose to conduct usage studies to more formally investigate how a robotic desk lamp could be used in a household context, thus informing the next generation design of such a robot. A slimmer, and safer, body layout is also called for.

12.5 Robotic theater acting

Part of the motivation for this work was to enable fluent and rehearsable human-robot theatrical performance. Much of this endeavor is left for future work.

We have staged a first such performance on the MIT campus in May 2007. In this performance, AUR, the robotic desk lamp described in Chapter 9 performed in a character part with two actors. However, the robot was not autonomous, but controlled using a hybrid control interface, which included scene loading, beat triggering, eye-contact control, and parametric control of joints on top of the scene animations.

We believe that this pilot play, as well as the concepts explored in this thesis, lay the groundwork to develop the human-robot theater project towards the aim of a fully autonomous robotic actor, one that can practice a scene with a human towards a fluently meshed performance.

12.6 Lessons for psychology research

It would be interesting to evaluate how joint practice of the type described in this work affects the mirror mechanisms described in Section 2.1. This could serve the evaluation of our claim that the proposed robotic cognitive architecture increasingly evokes a "like-me" sensation in the human teammate, one that is significantly stronger than in a purely reactive robot.

Finally, finding a principled way of using our infrastructure to inform psychological research, and in particular the nascent field of human-human joint action, would be a worthwhile effort.

Appendices

Appendix A

Fluency Study Questionnaire

This appendix lists the complete list of questions administered in the post-survey questionnaire described in Section 7.5.

The questionnaire included 41 questions. 38 questions asked the subjects to rank agreement with a sentence on a 7-point Likert scale, with the endpoints and midpoint labeled "Strongly Disagree" (1), "Neutral" (4), and "Strongly agree" (7). Three questions were open ended responses. It was administered immediately after completion of the experiment, on-screen via a web-based survey at a computer in our laboratory.

A.1 Questionnaire

1. My overall experience was enjoyable.
2. My overall experience was boring.
3. I would like to repeat the task again.
4. I believe the robot was controlled by a human behind the scenes.
5. The human-robot team did well on the task.
6. The human-robot team improved over time.

7. The human-robot team worked fluently together
8. The human-robot team's fluency improved over time.
9. The human-robot team felt well-tuned.
10. The robot did its part successfully.
11. The robot's performance improved over time.
12. The robot contributed to the fluency of the interaction.
13. The robot's performance was an important contribution to the success of the team.
14. I trusted the robot to do the right thing at the right time.
15. The robot was intelligent.
16. The robot was trustworthy.
17. The robot was committed to the task.
18. The robot performed well as part of the team.
19. Open-Ended Response: In a few sentences, please add any other comments about the robot's performance.
20. I learned to adapt my actions to the robot's actions.
21. The robot learned to adapt its actions to mine.
22. I had to carry the weight to make the human-robot team better.
23. The robot contributed equally to the team performance.
24. I was the most important team member on the team.
25. The robot was the most important team member on the team.
26. It felt like the robot was committed to the success of the team.
27. I was committed to the success of the team.

28. I felt increasing pressure to perform well as the team's performance improved.

29. Open-Ended Response: In a few sentences, please add any other comments about the team's performance.

30. I feel uncomfortable with the robot.

31. The robot and I understand each other.

32. The robot perceives accurately what my goals are.

33. I find what I am doing with the robot confusing.

34. I believe the robot likes me.

35. The robot does not understand what I am trying to accomplish.

36. The robot and I respect each other.

37. I am confident in the robot's ability to help me.

38. The robot and I are working towards mutually agreed upon goals.

39. I feel that the robot appreciates me.

40. The robot and I trust each other.

41. Open-Ended Response: If there is anything else you'd like to let us know about your experience, please write it below.

A.2 Statistical Metrics

This section lists statistical metrics, including significance in different between the two conditions, for all questions in the survey. All values are on a 7-point Likert scale reported as $mean \pm s.d.$.

1. **"My overall experience was enjoyable."**

   REACTIVE: 5.73 ± 1.03

   FLUENCY: 5.83 ± 1.1

   t(31)=-0.268 not signif.

2. **"My overall experience was boring."**

   REACTIVE: 2.87 ± 1.64

   FLUENCY: 2.39 ± 1.75

   t(31)=0.802 not signif.

3. **"I would like to repeat the task again."**

   REACTIVE: 4.53 ± 1.64

   FLUENCY: 4.61 ± 1.72

   t(31)=-0.132 not signif.

4. **"I believe the robot was controlled by a human behind the scenes."**

   REACTIVE: 1.8 ± 0.94

   FLUENCY: 2.5 ± 1.95

   t(31)=-1.271 not signif.

5. **"The human-robot team did well on the task."**

   REACTIVE: 5.47 ± 0.83

   FLUENCY: 5.22 ± 1.26

   t(31)=0.641 not signif.

6. **"The human-robot team improved over time."**

   REACTIVE: 5.87 ± 0.83

   FLUENCY: 6.17 ± 1.29

   t(31)=-0.773 not signif.

7. **"The human-robot team worked fluently together"**

   REACTIVE: 4.87 ± 0.92

   FLUENCY: 5.61 ± 0.98

   t(31)=-2.24 **$p < 0.05$** *

8. **"The human-robot team's fluency improved over time."**

   REACTIVE: 5.33 ± 1.4

   FLUENCY: 6.06 ± 1.3

   t(31)=-1.533 not signif.

9. **"The human-robot team felt well-tuned."**

   REACTIVE: 4.87 ± 1.51

   FLUENCY: 5.44 ± 0.86

   t(31)=-1.384 not signif.

10. **"The robot did its part successfully."**

    REACTIVE: 5.6 ± 0.51

    FLUENCY: 6.11 ± 1.02

    t(31)=-1.761 not signif.

11. **"The robot's performance improved over time."**

    REACTIVE: 4.27 ± 1.44

    FLUENCY: 6.28 ± 1.02

    t(31)=-4.695 **$p < 0.001$** ***

12. **"The robot contributed to the fluency of the interaction."**

    REACTIVE: 4.73 ± 1.22

    FLUENCY: 6.11 ± 1.18

    t(31)=-3.282 **$p < 0.01$** **

13. **"The robot's performance was an important contribution to the success of the team."**

    REACTIVE: 5.2 ± 1.37

    FLUENCY: 6.5 ± 0.62

    t(31)=-3.61 **$p < 0.01$** **

14. **"I trusted the robot to do the right thing at the right time."**

    REACTIVE: 4.93 ± 1.62

    FLUENCY: 5.44 ± 1.38

    t(31)=-0.977 not signif.

15. **"The robot was intelligent."**

    REACTIVE: 4.2 ± 1.7

    FLUENCY: 5.33 ± 1.08

    t(31)=-2.322 **$p < 0.05$** *

16. **"The robot was trustworthy."**

    REACTIVE: 4.87 ± 1.3

    FLUENCY: 5.39 ± 1.33

    t(31)=-1.132 not signif.

17. **"The robot was committed to the task."**

    REACTIVE: 5.53 ± 1.6

    FLUENCY: 5.5 ± 1.54

    t(31)=0.061 not signif.

18. **"The robot performed well as part of the team."**

    REACTIVE: 5.67 ± 0.98

    FLUENCY: 5.83 ± 1.2

    t(31)=-0.432 not signif.

19. **Open-ended question**

20. **"I learned to adapt my actions to the robot's actions."**

    REACTIVE: 6.0 ± 0.93

    FLUENCY: 5.72 ± 1.53

    t(31)=0.616 not signif.

21. **"The robot learned to adapt its actions to mine."**

    REACTIVE: 3.47 ± 1.46

    FLUENCY: 5.94 ± 1.06

    t(31)=-5.656 **$p < 0.001$** ***

22. **"I had to carry the weight to make the human-robot team better."**

    REACTIVE: 5.53 ± 1.25

    FLUENCY: 4.0 ± 1.78

    t(31)=2.806 **$p < 0.01$** **

23. **"The robot contributed equally to the team performance."**

REACTIVE: 3.93 ± 1.62

FLUENCY: 4.61 ± 1.33

t(31)=-1.317 not signif.

24. **"I was the most important team member on the team."**

REACTIVE: 5.33 ± 1.88

FLUENCY: 3.89 ± 2.0

t(31)=2.126 **$p < 0.05$** *

25. **"The robot was the most important team member on the team."**

REACTIVE: 2.33 ± 1.18

FLUENCY: 3.28 ± 1.71

t(31)=-1.811 not signif.

26. **"It felt like the robot was committed to the success of the team."**

REACTIVE: 4.47 ± 2.13

FLUENCY: 5.22 ± 1.56

t(31)=-1.175 not signif.

27. **"I was committed to the success of the team."**

REACTIVE: 6.4 ± 0.74

FLUENCY: 5.83 ± 1.1

t(31)=1.702 not signif.

28. **"I felt increasing pressure to perform well as the team's performance improved."**

REACTIVE: 6.33 ± 0.72

FLUENCY: 5.89 ± 1.23

t(31)=1.23 not signif.

29. **Open-ended question**

30. **"I feel uncomfortable with the robot."**

REACTIVE: 2.67 ± 1.59

FLUENCY: 2.67 ± 1.71

t(31)=0.0 not signif.

31. **"The robot and I understand each other."**

REACTIVE: 4.0 ± 1.77

FLUENCY: 4.44 ± 1.2

t(31)=-0.856 not signif.

32. **"The robot perceives accurately what my goals are."**

REACTIVE: 3.27 ± 1.62

FLUENCY: 4.61 ± 1.58

t(31)=-2.406 **$p < 0.05$** *

33. **"I find what I am doing with the robot confusing."**

REACTIVE: 2.6 ± 1.59

FLUENCY: 2.72 ± 1.6

t(31)=-0.219 not signif.

34. **"I believe the robot likes me."**

REACTIVE: 3.67 ± 1.76

FLUENCY: 3.44 ± 1.89

t(31)=0.347 not signif.

35. **"The robot does not understand what I am trying to accomplish."**

REACTIVE: 4.0 ± 1.93

FLUENCY: 3.17 ± 1.62

t(31)=1.351 not signif.

36. **"The robot and I respect each other."**

REACTIVE: 4.2 ± 1.42

FLUENCY: 4.44 ± 1.5

t(31)=-0.476 not signif.

37. **"I am confident in the robot's ability to help me."**

REACTIVE: 5.07 ± 0.88

FLUENCY: 5.17 ± 1.2

t(31)=-0.268 not signif.

38. **"The robot and I are working towards mutually agreed upon goals."**
    `REACTIVE:` 3.67 ± 1.72
    `FLUENCY:` 4.67 ± 1.61
    t(31)=-1.724 not signif.

39. **"I feel that the robot appreciates me."**
    `REACTIVE:` 3.13 ± 1.64
    `FLUENCY:` 3.44 ± 1.62
    t(31)=-0.547 not signif.

40. **"The robot and I trust each other."**
    `REACTIVE:` 3.8 ± 1.47
    `FLUENCY:` 4.28 ± 1.41
    t(31)=-0.951 not signif.

41. **Open-ended question**

A.3 Statistical Differences by Gender

The following lists statistical metrics, including significance in difference between male and female subjects, for those questions in the survey in which these two groups differed significantly. All values are on a 7-point Likert scale reported as *mean* ± *s.d.*.

- **"The human-robot team did well on the task."**
  `Male:` 5.71 ± 0.92
  `Female:` 4.94 ± 1.12
  t(31)=2.156 **$p < 0.05$** *

- **"The robot was intelligent."**
  `Male:` 4.24 ± 1.68
  `Female:` 5.44 ± 0.96
  t(31)=-2.502 **$p < 0.05$** *

- **"The robot contributed equally to the team performance."**

  Male: 3.76 ± 1.3

  Female: 4.88 ± 1.5

  t(31)=-2.276 **$p < 0.05$** *

- **"I was the most important team member on the team."**

  Male: 5.53 ± 1.23

  Female: 3.5 ± 2.25

  t(31)=3.24 **$p < 0.01$** **

- **"It felt like the robot was committed to the success of the team."**

  Male: 4.12 ± 2.0

  Female: 5.69 ± 1.3

  t(31)=-2.657 **$p < 0.05$** *

- **"The robot perceives accurately what my goals are."**

  Male: 3.29 ± 1.76

  Female: 4.75 ± 1.34

  t(31)=-2.66 **$p < 0.05$** *

Appendix B

Fluency Experiment Instructions

These were the instructions given to subjects in the Fluency experiment described in Chapter 6. They were identical for all subjects, regardless of experimental condition.

In addition, subjects were verbally reminded of safety concerns, encouraged to stay clear of the robot's motion path, and were told that if they knew what color is written behind a door, there is no imperative to open the door as long as they can make the robot shine the correct light on the board. They were also verbally reminded that the robot can only go to their hand location if their hand is still.

B.1 Instruction Text

Robotic Desk Lamp Experiments

INSTRUCTIONS

Thank you for participating in the Robotic Desk Lamp Experiment. Please read these instructions carefully and ask the experimenter if you have any questions.

Safety

Please be aware that the robot can move relatively fast and is very strong. Make sure that you keep your hands and head at a safe distance from the robot at all times. Since the robot might make unexpected moves, make sure that you have a good sense where the robot is at all times and be alert to the sound of its motion.

The Task

In this study you play a repetition game in collaboration with a robot. You will play two separate rounds of the game which will be scored independently.

Each task is made up of eight (8) steps, which will be given to you by the experimenter. Each step is denoted by a letter (A-C) and a pattern:

For each step you need to go to the board indicated by the letter, and open the door indicated by the pattern. Please open the door carefully without ripping the cardboard. You may use the provided tool to open the door.

Behind each door you will find the name of a color. The goal of each step is to make the lamp shine the correct color on the indicated board. Make sure the door is completely closed when the light is pointed towards the board. When this is accomplished, the experimenter will sound a single tone to indicate that you can move to the next step.

When all eight steps are completed a double tone will be sounded, indicating you have completed the sequence. The experimenter will inform you of your total sequence time.

You are required to repeat each sequence ten (10) times.

Controlling the robot

You control the robot using voice commands and your hand. In order to do so you will be equipped with a motion-detection glove and a microphone. The robot understands the following commands:

- "Go" / "Come" / "Come Here" These three commands are equivalent and will cause the robot to go where your hand is at the moment. Note that the robot cannot detect your hand position while your hand is moving.
- "Red" / "Green" / "Blue" / "White" / "Off" These five commands will cause the light to turn into the indicated color or turn off.

The robot might move independent of your commands. This is normal behavior, but does not necessarily occur. If the robot is at the correct place and shining the correct light, the move will be counted as completed, even if you did not issue any voice commands.

Time

Your goal is to complete each sequence in the least amount of time. A sequence time is measured between two double-tone signals. The experimenter will tell you the time it took you to complete the sequence at the end of each sequence.

Completing the Study

You play the game until you've completed 10 repetitions of the first sequence, and 10 repetitions of the second sequence. At this point you can choose to do 5 more repetitions of a third sequence, or notify the experimenter that youre done.

You will then be asked to complete a questionnaire regarding your experience in the study.

Practice & Questions

Please take a moment to get used to how you control the robot and how you open the board doors. You can practice until you feel comfortable with how the game operates. Then tell the experimenter that you are ready to go.

Thank you and good luck!

Bibliography

R. Alami, A. Clodic, V. Montreuil, E. A. Sisbot, and R. Chatila. Task planning for human-robot interaction. In *sOc-EUSAI '05: Proceedings of the 2005 joint conference on Smart objects and ambient intelligence*, pages 81–85, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-304-2.

B. Argall, Y. Gu, B. Browning, and M. Veloso. The first segway soccer experience: towards peer-to-peer human-robot teams. In *HRI '06: Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 321–322, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-294-1.

B. Argall, B. Browning, and M. Veloso. Learning by demonstration with critique from a human teacher. In *HRI '07: Proceeding of the ACM/IEEE international conference on Human-robot interaction*, pages 57–64, New York, NY, USA, 2007. ACM Press. ISBN 978-1-59593-617-2.

D. Baldwin and J. Baird. Discerning intentions in dynamic human action. *Trends in Cognitive Sciences*, 5(4):171–178, 2001.

L. Barsalou. Private Communication, 2007.

L. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–660, 1999.

L. W. Barsalou, P. M. Niedenthal, A. Barbey, and J. Ruppert. Social embodiment. *The Psychology of Learning and Motivation*, 43:43–92, 2003.

R. Bischoff and V. Graefe. Hermes – a versatile personal assistant robot. *Proc. IEEE – Special Issue on Human Interactive Robots for Psychological Enrichment*, pages 1759–1779, 2004.

A. Boal. *Games for Actors and Non-Actors*. Routledge, 2nd edition, July 2002.

M. Bratman. Shared cooperative activity. *The Philosophical Review*, 101(2):327–341, 1992.

C. Breazeal. *Designing Sociable Robots*. MIT Press, 2002. ISBN 0262025108.

C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Edmonton, Alberta, Canada, August 2005.

C. Breazeal, A. Wang, and R. Picard. Experiments with a robotic computer: body, affect and cognition interactions. In *HRI '07: Proceeding of the ACM/IEEE international conference on Human-robot interaction*, pages 153–160, New York, NY, USA, 2007. ACM Press. ISBN 978-1-59593-617-2.

C. Bregler. Learning and recognizing human dynamics in video sequences. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, page 568. IEEE Computer Society, 1997. ISBN 0-8186-7822-4.

S. Broadhurst. The Jeremiah Project: Interaction, Reaction, and Performance. *The Drama Review*, 48(4), 2004.

R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

A. Bruce, J. Knight, S. Listopad, B. Magerko, and I. Nourbakhsh. Robot improv: Using drama to create believable agents. In *Proceedings of ICRA 2000*, volume 4, pages 4002–4008, April 2000.

A. Bruce, I. Nourbakhsh, and R. Simmons. The role of expressivness and attention in human-robot interaction. In *Proceedings of the 2001 AAAI Fall Symposium*, Cape Cod, MA, USA, 2001. AAAI Press.

D. J. Bruemmer, D. D. Dudenhoeffer, and J. Marble. Dynamic autonomy for urban search and rescue. In *2002 AAAI Mobile Robot Workshop*, Edmonton, Canada, August 2002.

R. Burke, D. Isla, M. Downie, Y. Ivanov, and B. Blumberg. CreatureSmarts: The art and architecture of a virtual brain. In *Proceedings of the Game Developers Conference*, pages 147–166, 2001.

N. Chovil. Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25:163–194, 1992.

H. H. Clark. *Using Language*. Cambridge University Press, Cambridge, UK, 1996.

P. R. Cohen and H. J. Levesque. Teamwork. *NOÛS*, 35:487–512, 1991.

T. Cole and H. K. Chinoy. *Actors on Acting : The Theories, Techniques, and Practices of the World's Great Actors, Told in Their Own Words*. Crown, 3th edition, 1970.

R. Cooper and T. Shallice. Contention scheduling and the control of routing activities. *Cognitive Neuropsychology*, 17:297–338, 2000.

A. R. Damasio. Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33:25–62, 1989.

D. C. Dennett. Three kinds of intentional psychology. In *The Intentional Stance*, chapter 3. MIT Press, Cambridge, MA, 1987.

S. Dixon. Metal Performance: Humanizing Robots, Returning to Nature, and Camping About. *The Drama Review*, 48(4), 2004.

F. Doshi and N. Roy. Efficient model learning for dialog management. In *HRI '07: Proceeding of the ACM/IEEE international conference on Human-robot interaction*, pages 65–72, New York, NY, USA, 2007. ACM Press. ISBN 978-1-59593-617-2.

B. R. Duffy. Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42 (3-4):177–190, 2003.

Y. Endo. Anticipatory and improvisational robot via recollection and exploitation of episodic memories. In *Proceedings of the AAAI Fall Symposium*, 2005.

J. R. Flanagan and R. S. Johansson. Action plans used in action observation. *Nature*, 424 (6950):769–771, August 2003. ISSN 1476-4687.

T. W. Fong, C. Thorpe, and C. Baur. Multi-robot remote driving with collaborative control. *IEEE Transactions on Industrial Electronics*, 2003.

V. Gallese and A. Goldman. Mirror neurons and the simulation theory of mind-reading. *Brain*, 2(12):493–501, 1996.

B. Gleissner, A. N. Meltzoff, and H. Bekkering. Children's coding of human action: cognitive factors influencing imitation in 3-year-olds. *Developmental Science*, 3(4):405–414, 2000.

M. Goodrich, D. Olsen, J. Crandall, and T. Palmer. Experiments in adjustable autonomy. In *Proceedings of the IJCAI Workshop on Autonomy, Delegation and Control: Interacting with Intelligent Agents*, 2001.

B. J. Grosz. Collaborative systems. *AI Magazine*, 17(2):67–85, 1996.

R. Hamdan, F. Heitz, and L. Thoraval. Gesture localization and recognition using probabilistic visual learning. In *Proceedings of the 1999 Conference on Computer Vision and Pattern Recognition (CVPR '99)*, pages 2098–2103, Ft. Collins, CO, USA, June 1999.

G. Hoffman. HRI: Four lessons from acting method. Technical report, MIT Media Laboratory, Cambridge, MA, USA, 2005a.

G. Hoffman. An experience-based framework for intention-reading. Technical report, MIT Media Laboratory, Cambridge, MA, USA, 2005b.

G. Hoffman and C. Breazeal. Collaboration in human-robot teams. In *Proc. of the AIAA 1st Intelligent Systems Technical Conference*, Chicago, IL, USA, September 2004. AIAA.

G. Hoffman and C. Breazeal. Cost-based anticipatory action-selection for human-robot fluency. *IEEE Transactions on Robotics and Automation (in press)*, 2007.

A. O. Horvath and L. S. Greenberg. Development and validation of the working alliance inventory. *Journal of Counseling Psychology*, 36(2):223–233, 1989.

C. Jones. The dot and the line. MGM Animation/Visual Arts, 1965.

H. Jones and S. Rock. Dialogue-based human-robot interaction for space construction teams. In *IEEE Aerospace Conference Proceedings*, volume 7, pages 3645–3653, 2002.

H. Kimura, T. Horiuchi, and K. Ikeuchi. Task-model based human robot cooperation using vision. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS'99)*, pages 701–706, 1999.

G. Knoblich and J. S. Jordan. Action coordination in groups and individuals: learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(5):1006–1016, September 2003. ISSN 0278-7393. doi: 10.1037/0278-7393.29.5.1006.

R. Kohavi, B. Becker, and D. Sommerfield. Improving simple bayes. In *Proceedings of the European Conference on Machine Learning*, 1997.

T. Komatsu and Y. Miyake. Temporal development of dual timing mechanism in synchronization tapping task. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Communication (RO-MAN 2004)*, September 2004.

S. Kosslyn. Mental imagery. In S. Kosslyn and D.N.Osherson, editors, *Invitation to Cognitive Science: Visual Cognition*, volume 2, chapter 7, pages 276–296. MIT Press, Cambridge, MA, 2nd edition, 1995.

R. M. Krauss, Y. Chen, and P. Chawla. Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In M. Zanna, editor, *Advances in experimental social psychology*, pages 389–450. Tampa: Academic Press, 1996.

G. Kreiman, C. Koch, and I. Fried. Imagery neurons in the human brain. *Nature*, 408: 357–361, 2000.

G. Lakoff and M. Johnson. *Philosophy in the flesh : the embodied mind and its challenge to Western thought*. Basic Books, New York, 1999.

J. Lasseter. Luxo jr. Pixar Animation Studios, 1986.

Les Freres Corbusier. Heddatron. http://www.lesfreres.org/heddatron/, 2006.

J. L. Locke and K. J. Kutz. Memory for speech and speech for memory. *Journal of Speech and Hearing Research*, 18:176–191, 1975.

B. Malle, L. Moses, and D. Baldwin, editors. *Intentions and Intentionality*. MIT Press, 2001.

A. Martin. Functional neuroimaging of semantic memory. In R. Cabeza and A. A. Kingstone Kingstone, editors, *Handbook of Functional Neuroimaging of Cognition*, pages 153–186. MIT Press, 2001.

E. Martinson and D. Brock. Improving human-robot interaction through adaptation to the auditory scene. In *HRI '07: Proceeding of the ACM/IEEE international conference on Human-robot interaction*, pages 113–120, New York, NY, USA, 2007. ACM Press. ISBN 978-1-59593-617-2.

D. Massaro and M. M. Cohen. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9(5):753–771, 1983.

M. J. Matarić. Sensory-motor primitives as a basis for imitation: linking perception to action and biology to robotics. In K. Dautenhahn and C. L. Nehaniv, editors, *Imitation in animals and artifacts*, chapter 15, pages 391–422. MIT Press, 2002. ISBN 0-262-04203-7.

E. McClave. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878, 2000.

P. McLeod and M. I. Posner. Privileged loops from percept to act. In H. Bouma and D. G. Bouwhuis, editors, *Attention and performance*, volume 10, pages 55–66. Erlbaum, Hillsdale, NJ, 1984.

S. Meisner and D. Longwell. *Sanford Meisner on Acting*. Vintage, 1st edition, August 1987.

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

M. P. Michalowski, S. Sabanovic, and H. Kozima. A dancing robot for rhythmic social interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 89–96, Arlington, Virginia, USA, March 2007.

S. Moore. *Training an Actor: The Stanislavski System in Class.* Viking Press, New York, NY, 1968.

L. M. Parsons. Temporal and kinematic properties of motor behavior reflected in mentally simulated action. *Journal of Experimental Psychology: Human Perception and Performance,* 20(4):709–730, Aug 1994.

D. Pecher and R. A. Zwaan, editors. *Grounding cognition: the role of perception and action in memory, language, and thinking.* Cambridge Univ. Press, Cambridge, UK, 2005.

A. Pentland. Social dynamics: Signals and behavior. In *Proceedings of the Third International Conference on Development and Learning: ICDL 2004,* La Jolla, CA, October 2004.

K. Perlin and A. Goldberg. Improv: a system for scripting interactive actors in virtual worlds. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques,* pages 205–216. ACM Press, 1996. ISBN 0-89791-746-4.

R. Pfeifer and J. C. Bongard. *How the Body Shapes the Way We Think: A New View of Intelligence (Bradford Books).* The MIT Press, 2007. ISBN 0262162393.

C. Rich, C. L. Sidner, and N. Lesh. Collagen: Applying collaborative discourse theory to human-computer collaboration. *AI Magazine,* 22(4):15–25, 2001.

D. K. Roy and A. P. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science,* 26(1):113–146, January 2002.

P. Rybski, K. Yoon, J. Stolarz, and M. Veloso. Interactive robot task training through dialog and demonstration. In *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI'07),* Washington DC, March 2007.

Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, and K. Fujimura. The intelligent asimo: System overview and integration. In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems,* volume 3, pages 2478–2483, 2002.

K. Sakita, K. Ogawam, S. Murakami, K. Kawamura, and K. Ikeuchi. Flexible cooperation between human and robot by interpreting human intention from gaze information. In

Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), pages 846–851, 2004.

J. R. Searle. Collective intentions and actions. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, chapter 19, pages 401–416. MIT Press, Camridge, MA, 1990.

N. Sebanz, H. Bekkering, and G. Knoblich. Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76, 2006.

C. Sidner, C. Lee, and N. Lesh. Engagement rules for human-robot collaborative interaction. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3957–3962, 2003.

C. L. Sidner, C. Lee, L.-P. Morency, and C. Forlines. The effect of head-nod recognition in human-robot conversation. In *HRI '06: Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 290–296, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-294-1.

K. Simmons and L. W. Barsalou. The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology*, 20:451–486, 2003.

K. O. Solomon and L. W. Barsalou. Representing properties locally. *Cognitive Psychology*, 43:129–169, 2001.

J. A. Somerville, A. L. Woodward, and A. Needham. Action experience alters 3-month-old infants' perception of others actions. *Cognition*, 2004.

J. Sonenberg. *The Actor Speaks : Twenty-four Actors Talk About Process and Technique*. Three Rivers Press, 1st edition, April 1996. Selected chapters.

M. J. Spivey, D. C. Richardson, and M. Gonzalez-Marquez. On the perceptual-motor and image-schematic infrastructure of language. In D. Pecher and R. A. Zwaan, editors, *Grounding cognition: the role of perception and action in memory, language, and thinking.* Cambridge Univ. Press, Cambridge, UK, 2005.

V. Spolin. *Improvisation for the Theater*. Northwestern University Press, Evanston, IL, USA, 3rd edition, 1999.

R. Stanfield and R. Zwaan. The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, 12:153–156, 2001.

G. Stebbins. *Delsarte System of Expression*. Edgar S. Werner, New York, NY, 2nd edition, 1887.

I. Thornton, P. J., and M. Shiffrar. The visual perception of human locomotion. *Cognitive Neuropsychology*, 15:535–552, 1998.

Various Authors. Pat-a-cake on wikipedia. `http://en.wikipedia.org/wiki/Pat-a-cake,_pat-a-cake,_baker's_man`, 2007.

W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfe. Sphinx-4: A flexible open source framework for speech recognition. Technical Report TR-2004-139, Sun Microsystems Laboratories, November 2004.

G. Weinberg and S. Driscoll. Robot-human interaction with an anthropomorphic percussionist. In *Proceedings of International ACM Computer Human Interaction Conference (CHI 2006)*, pages 1229–1232, Montréal, Canada, 2006.

M. Wilson. Perceiving imitable stimuli: consequences of isomorphism between input and output. *Psychological Bulletin*, 127(4):543–553, 2001.

M. Wilson. Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4):625–636, December 2002.

M. Wilson and G. Knoblich. The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131:460–473, 2005.

S. Wilson, A. Saygin, M. Sereno, and M. Iacoboni. Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, 7:701–702, 2004.

H. Woern and T. Laengle. Cooperation between human beings and robot systems in an industrial environment. In *Proceedings of the Mechatronics and Robotics*, volume 1, pages 156–165, 2000.

A. L. Woodward, J. A. Sommerville, and J. J. Guajardo. How infants make sense of intentional actions. In B. Malle, L. Moses, and D. Baldwin, editors, *Intentions and Intentionality: Foundations of Social Cognition*, chapter 7, pages 149–169. MIT Press, Cambridge, MA, 2001.

C. Wren, B. Clarkson, and A. Pentland. Understanding purposeful human motion. In *Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 378–383, 2000.

C. R. Wren and A. Pentland. Dynaman: A recursive model of human motion. Technical Report VisMod 451, MIT Media Laboratory, 1997.