Intelligent Information Sharing to Support Loosely-Coupled Teamwork

a dissertation presented

by

Ofra Amir

to

The School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Computer Science

Harvard University

Cambridge, Massachusetts

November, 2016

Dissertation advisor: Barbara J. Grosz Ofra Amir

Intelligent Information Sharing to Support Loosely-Coupled Teamwork

Abstract

Complex tasks such as treating patients with complex medical conditions, conducting research, co-authoring documents and developing software products are typically accomplished by teams. Teamwork in such settings is often loosely-coupled as team members assume different responsibilities that match their individual expertise. This decomposition of activities enables team members to function autonomously and requires team members to be aware of others' actions only if these actions interact with their own activities. However, identifying interactions between collaborators' activities can be challenging. As a result, team members are often overwhelmed by too much irrelevant information about others' activities, or lack important relevant information, both of which can lead to coordination failures. This thesis argues that intelligent information sharing methods that identify the information that is most relevant to each team member can reduce coordination overhead and improve team performance.

Through a study of teams caring for children with complex medical conditions, this thesis characterizes the coordination challenges of loosely-coupled teams and formalizes the problem of information sharing in such teams. The thesis introduces Mutual Influence Potential Networks, a new representation for modeling collaborative activities. It further defines MIP-DOI, an algorithm which uses the Mutual Influence Potential Network representation to identify the most relevant information for each team member.

The thesis also presents the design, implementation and evaluation of a personalized change awareness mechanism, which uses MIP-DOI to reduce the amount of shared change information in the context of collaborative writing. The results of an experiment evaluating

this mechanism show that compared to the currently most prevalent approach of presenting users with all changes made by their collaborators, the personalized change awareness mechanism resulted in significantly reduced perceived workload and significantly increased productivity of team members. Importantly, the personalized change awareness mechanism did not have any detrimental effect on the quality of the work.

Contents

Citations to Previously Published Work

The work presented in Chapter 2 was done with our clinical collaborators at Stanford's Lucile Packard Children's Hospital, who provided medical expertise and helped design and conduct the interviews. Significant portions of this research have appeared in the following paper:

Ofra Amir, Barbara Grosz J., Krzysztof Z. Gajos, Sonja Swenson, and Lee Sanders. From care plans to care coordination: Opportunities for computer support of teamwork in complex healthcare. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, pages 1419–1428, 2015.

Significant portions of Chapter 3 have appeared in the following paper:

Ofra Amir, Barbara Grosz J., and Krzysztof Z. Gajos. Mutual influence potential networks: Enabling information sharing in loosely-coupled extended-duration teamwork. In *Proceedings of the Twenty-Fifth international joint conference on Artificial Intelligence (IJCAI'16)*, pages 796–803, 2016.

The research presented in Chapter 6 was done during an internship at Microsoft Research. It has appeared in the following paper:

Ofra Amir, Ece Kamar, Andrey Kolobov, and Barbara J. Grosz. Interactive teaching strategies for agent training. newblock In *Proceedings of the Twenty-Fifth international joint conference on Artificial Intelligence (IJCAI'16)*, pages 804–811, 2016.

Acknowledgments

First and foremost, I wish to thank my advisor Barbara Grosz. Barbara has won nearly every award that exists, but what continues to impress me most is seeing how much she cares about her students and colleagues and the effort she makes to help them achieve their goals. I thank Barbara for her mentorship, advice, support and insights, for being patient with my occasional stubbornness and for reading everything I wrote (except these acknowledgments) more times than any other person would have been willing to.

Krzysztof Gajos contributed tremendously to this thesis. Krzysztof taught me to think more seriously about people when — or rather *before* — thinking about AI solutions. His feedback typically suggested a different way of thinking about a problem rather than a particular solution, and as such it always led me in interesting directions. I also thank Krzysztof for adopting me to his group when I was Barbara's only student, and for tea breaks.

Perhaps unsurprisingly for a thesis concerned with teams, it was the result of productive teamwork. I cannot think of a better way to conduct research at the intersection of AI and HCI than trying to integrate Barbara's and Krzysztof's (often conflicting) advice. Watching them misinterpret each other in brilliant ways was inspiring and entertaining at once. I also had the pleasure of TF-ing with Barbara and Krzysztof and learned a lot from them in that respect, too. I feel truly privileged to have both Barbara and Krzysztof on my corner, and am proud to say that the three of us make a pretty good team.

While we did not solve all care coordination problems quite yet, the ideas in this thesis were largely inspired by what I learned the complex care environment. I am grateful to my collaborators at Stanford, Lee Sanders, Sonja Swenson and Jody Lin for educating me about their world and for joining us on this research adventure. Many thanks also go to the families and providers who were willing to share their experiences. I thank Stan Rosenschein for taking interest in my work and for many helpful discussions during my Stanford visits.

Talking to Stan about what I learned in real-time helped me organize my thoughts, and Stan's thoughtful observations contributed a lot to my thinking about teamwork and coordination. I thank Ece Kamar and Andrey Kolobov for a fruitful internship at MSR which resulted in the work described in Chapter 6. I thank my student collaborators: Logan Martin and Neel Patel helped develop GoalKeeper; Lauren Urke, Sebastian Gehrmann, Ezra Zigmond and Limor Gultchin contributed to the work on collaborative writing.

David Parkes gave me great advice along the way and has always been positive and encouraging. As the CS area dean, David also contributed a lot to making Harvard CS a really great environment to work in. Thanks also to other CS faculty members Stuart, Radhika, Yiling, Margo, Finale and Yaron for helpful research conversations.

Thanks to Barbara's group members Edith, Roni, Sebastian, Shiri, Avshalom and Fei and to the IIS group members for providing a stimulating and friendly research environment, to Peter Krafft for initiating and co-organizing the Harvard-MIT teamwork reading group, and to Yuval Hart for fun discussions about research, the universe and everything.

A special thanks goes to my first research mentor Kobi Gal. Kobi not only triggered my desire to study abroad and helped me achieve it, but also helped make the transition very smooth (he even made sure I got an appropriate winter coat). Thanks to Dave Rand who was the first to open Harvard's doors to me for a great internship at the Berkman Center and for teaching me about the parallels between doing research and making music.

Thanks to my office-mates over the years in MD 240: Katharina, for being a great researcher role model and more importantly for her friendship and Burdick's excursions; Roni, for long brainstorming sessions; James, for basketball breaks; part-time officemates Finale and Liz, for fun discussions; Jae, for making staying at the office late fun, and to my most recent office-mates Ken, Anna, Pao, Steve, Bernd and Sebastian, for making our office a place where no opportunity for procrastination is missed. Thanks also to fellow G5 students Michael Crouse and Sam Wiseman for commiserating since our first orientation

session. Thanks to admins David Lopez and Jess Jackson for their help with bureaucracy.

As it turns out, a PhD is often a test to one's perseverance and confidence. In that respect, I thank my teenage years mentors in "Hugey Sayarut" who taught me to cope with unexpected obstacles in the outdoors. I am especially thankful to Elli Booch, whose vote of confidence in me left me no choice but to eventually become confident.

The Israeli gang contributed to making life in Cambridge fun. Thanks to Rotem & Ronni for sharing the adventure from day one, to Omer for roommating, hot chocolate breaks and ramen, and to Yuval & Netta although they left way too early. Thanks to Yael and to my cousins Roy and Gali and their families for hosting me during my Stanford excursions. Thanks to my friends in Israel who were willing to hangout even when I didn't deliver Amazon packages, especially to Mey-Tal, Efrat, Assaf, Erez & Tali, Ofer, Aviad & Hadar and Itamar.

I am fortunate to have, in addition to my family back home, a family here in Boston. A very special thanks goes to Hanna, Ami, Yonatan, Alon and Noga, who accepted me as a Levy-Moonshine although I'm an Amir, and made Boston feel like home from day one.

And finally, a big thanks goes to my family back home: mom, dad, Dror, Dan and Shira. I thank my parents for letting me make decisions on my own since I was a toddler. I thank my siblings for making fun of me and keeping me sharp when I visited home. Also thanks to younger members of the extended family, Mila, Na'ama and Raz, who were always willing to play with me even though I consistently missed their birthdays.

This thesis work was supported in part by NIH grant 1R01CA204585-01 as part of the NSF/NIH Smart and Connected Health program, the Nuance Foundation, a Siebel scholarship, a CIMIT student technology in healthcare prize and a Harvard Mind, Brain and Behavior graduate student award.

Chapter 1

Introduction

Teamwork is a core human activity, essential to progress in many areas. A vast body of research in the social sciences, human-computer interaction (HCI) and computer-supported cooperative work (CSCW) has studied teamwork and developed tools to support teamwork. These tools and insights have enabled teams to work together more effectively in many settings, but there remain settings in which teams fail to coordinate, and for which technology does not yet provide adequate support. In particular, highly distributed teams with loosely coupled activities face significant coordination challenges.

The care of children with complex medical conditions, one of the domains this thesis studies, is a notable example of such settings. These children require care from a large, diverse set of caregivers that includes parents and community support organizations and multiple types of medical professionals. Despite widespread consensus on a need for greater coordination among providers, care for most of these children remains poorly coordinated [Leape, 2012]. As a result, they have high rates of unmet health needs, suboptimal physical functioning, and potentially preventable health care crises, and they account for a disproportionate share of health-system use and costs [Leape, 2012].

To ensure coordination, team members need to be aware of others' activities if they

interact with their own activities. For example, in healthcare teams each of the care providers is responsible for a different aspect of the patient's treatment, but providers need to know of others' activities if they interact with their own activities. For instance, a physical therapist would need to know about a drug prescribed by the neurologist if it affected the mobility of the patient. Identifying when the activities of team members interact can be particularly difficult in loosely-coupled teamwork. Consequently, team members often face one or both of the following coordination challenges: (1) high coordination overhead as a result of information overload when too much information is shared, or (2) coordination failure due to lack of important information when too little information is shared or when relevant information cannot be found.

This thesis argues that the information shared with a team member about collaborative activities should be personalized, such that only the subset of information that is most relevant to that team member will be presented. Such personalized sharing of information has the potential to improve coordination in distributed loosely-coupled teamwork through simultaneously (1) lowering coordination overhead by reducing the total amount of change information each team member needs to review, and (2) improving the chances of coordination success by ensuring that team members receive the information most relevant to their own tasks. The thesis of this dissertation is thus that:

Intelligent information sharing mechanisms that identify and share with team members relevant information about others' activities can reduce coordination overhead and improve team performance.

The dissertation presents a formative study of complex healthcare teams which uncovered and characterized new challenges in the coordination of loosely-coupled teamwork. It presents a new representation and an algorithm that uses this representation to reason about information sharing in such teamwork. It also presents an implementation of a personalized change awareness mechanism that uses the information sharing algorithm to share with each

team member the subset of information deemed most relevant for their own activities. It presents an evaluation of the personalized change awareness mechanism, demonstrating that personalized information sharing can reduce team members' perceived workload and increase their productivity, without causing any detrimental effects to work quality.

1.1 FLECS Teamwork

Healthcare research has shown that effective systems of care for patients with complex chronic conditions require an *integrated care plan* that addresses patient-centered health goals and provides context for treatment decisions for all members of the care team [McAllister, 2014, Adams et al., 2014]. To better understand the barriers to forming and using such care plans, we conducted a study of teams complex care teams. We interviewed and observed representative team members in the Complex Primary Care Clinic (*CPCC*) at Stanford's Lucile Packard Children's Hospital, including parents of children with complex conditions, primary care providers (PCP), medical specialists, therapists and administrators.

The study revealed five characteristics of complex care that raise significant challenges to effective teamwork:

- ***F****lat-structure of team with consensus-driven plan development.*
- ***L****oosely coupled plans of individual team members.*
- ***E****xtended duration of plans.*
- ***C****ontinual distributed revision of plans.*
- ***S****yncopated time scales of different team members.*

Taken together, these teamwork characteristics, denoted FLECS, result in a teamwork that differs fundamentally from the teamwork settings addressed by prior work in the social sciences and computer supported cooperative work (*CSCW*). To derive design implications

for developing systems to support FLECS teamwork, we used SharedPlans (SP) [Grosz and Kraus, 1996], a computational theory of collaboration which provides a specification of teamwork general enough to cover such teams. A SharedPlans-based analysis suggested the following key roles for technology for supporting FLECS teamwork: (1) making the team's plan "ever present", adapting the content and form of its presentation to individual team members based on their involvement in the plan and context of use; (2) supporting efficient information sharing by team members, and (3) enabling team members to easily adapt and expand parts of the plan, while ensuring their changes do not conflict with others' activities.

1.2 Mutual Influence Potential Networks: Reasoning About Information Sharing

One of the key opportunities for technology support of teamwork identified by our study of complex care teams was supporting efficient information sharing with team members. We formally define the problem of Information Sharing in Loosely-Coupled Teamwork (ISLET) and present new methods for addressing it. To support team coordination, solutions to the ISLET problem need to identify and share with team members information that is *relevant* to their activities under a *limited communication budget* so as to not overwhelm them with too much information.

Prior work in multi-agent systems developed algorithms to reason about information sharing in *agent* teams [Roth et al., 2006, Amir et al., 2014, Melo et al., 2012, Wu et al., 2011, Unhelkar and Shah, 2016]. However, these approaches rely on a *complete plan knowledge assumption*, which does not hold in many *human* teamwork settings. For example, members of the complex health care teams we studied might agree on high-level treatment goals but never fully specify a long-term plan. The approach presented in this thesis does not rely on the availability of a complete model of the team's plan. Instead, it utilizes the extended duration

of teamwork to learn collaboration patterns from team members' interactions with each other and with different tasks. We introduce a new representation, "Mutual Influence Potential Network" (MIP-Net), to model knowledge about such collaboration patterns. MIP-Nets are updated over time based on the system's observations of team member interactions. MIP-Nets implicitly represent role allocation (i.e., team members' primary responsibilities) and dependencies between different team members' activities. We defined the MIP-DOI algorithm which uses the MIP-Net structure to reason about information sharing decisions. An empirical evaluation of this approach using a simulation of collaborative activities demonstrated that MIP-DOI can identify relevant information to share with team members.

1.3 Personalized Change Awareness: Reducing Information Overload

We designed and implemented a *personalized change awareness* mechanism that uses MIP-Nets and MIP-DOI to support collaborative writing. *Change awareness* mechanisms [Tam and Greenberg, 2006, Dourish and Bellotti, 1992] such as change tracking features and Diff tools are a prominent approach for supporting coordination in distributed teams. These mechanisms assist team members in tracking each others' activities. By enhancing team members' knowledge of others' activities, they provide context for evaluating their own actions and ensuring they align with the team's activity as a whole [Dourish and Bellotti, 1992].

While some change awareness mechanisms allow team members to filter the information presented to them (e.g., to omit styling changes in a document, to show changes made by a particular collaborator), they do not reason about the *relevance* of such information to a particular team member. In contrast, we developed a personalized change awareness mechanism for supporting collaborative writing that *automatically* selects the changes (document edits) to share with each team member using the MIP-DOI algorithm. When an author returns to

a shared document, the mechanism highlights a subset of edits made by collaborators that are deemed most relevant to her, thus limiting the amount of information she needs to review prior to making her own writing contributions.

We conducted an experiment which compared personalized change awareness to two baselines: a change awareness mechanism that showed *all* of the changes, and a change awareness mechanism that showed the same number of changes as the personalized mechanism, but selected the changes to share at *random*. Our results showed that both mechanisms that restricted the number of changes shown, *Personalized* and *Random*, led to reduced participants' perceived workload and higher productivity compared to the condition in which all changes were shared, demonstrating the benefits of personalized change awareness with respect to reducing coordination overhead. They also showed that personalized changes shared with participants were rated as more helpful than randomly chosen changes, and that the quality of the team's documents was higher with the personalized mechanism compared to the random mechanism, demonstrating the importance of sharing relevant changes.

1.4 GoalKeeper: Supporting Complex Care Teams

Our study of complex care teams revealed several barriers to the implementation of integrated team-baed care plans, one of which is the lack of appropriate technology support for managing care plans. While electronic medical records (EMR) have been shown to facilitate improved coordination within health organizations, they were found ineffective for coordinating providers in different organizations and were also found to create information overload [O'Malley et al., 2010]. They support linear processes rather than dynamic coordination processes and are optimized for billing uses rather than for provider and patient needs [O'Malley et al., 2010]. We developed GoalKeeper, a system for supporting the creation and management of team-based care plans. GoalKeeper enables families and their care

providers to define care goals for a child, specify actions that need to be taken to make progress toward these goals and track progress toward goals by continuously recording status updates about the child's condition.

We conducted a preliminary pilot study of GoalKeeper, in which it was used by two families to create and track care plans. Study participants found GoalKeeper helpful for organizing their activities and prompting them to act when they accomplished goals or when they were not able to make progress. However, the study also revealed challenges in consistently engaging with GoalKeeper and a negative emotional effect on families when they observed lack of progress toward goals.

1.5 Interactive Teaching Strategies for Student-Teacher Reinforcement Learning

Agents learning to act in new environments can benefit from the advice of more experienced agents or from human teachers. In the student-teacher reinforcement learning framework, an experienced "teacher" agent helps accelerate the "student" agent's learning by providing advice on which action to take next [Clouse, 1996, Torrey and Taylor, 2013]. We developed new interactive teaching strategies for this framework.

In contrast to prior approaches, the teaching strategies we developed involve both the student and the teacher in the decision-making process. In these *jointly-initiated* teaching strategies, the student determines whether to ask for the teacher's attention, and the teacher, if asked to pay attention to the student's state, decides whether to use this opportunity to give advice, given a limited advice budget. An empirical evaluation demonstrated that the jointly-initiated teaching strategies reduced the amount of attention required of the teacher compared to teacher-initiated strategies, while maintaining similar learning gains.

1.6 Contributions and Thesis Overview

The major contributions of this thesis are as follows:

- Through a study of complex healthcare teams, it characterizes challenges and opportunities for technology support of FLECS teamwork. (Chapter 2)
- It develops a new representation, mutual influence potential networks (MIP-Nets) to model collaborative activities, and a new algorithm, MIP-DOI, that uses MIP-Nets to reason about the problem of information sharing in loosely-coupled extended-duration teamwork. It demonstrates the ability of MIP-DOI to identify relevant information in a collaborative activity simulation. (Chapter 3)
- It develops a personalized change awareness mechanism for supporting collaborative writing. This mechanism uses MIP-DOI to limit the amount of change information that is shared with each team member. It presents an evaluation of this mechanism, demonstrating the benefits of personalized change awareness to teamwork. (Chapter 4)
- It presents GoalKeeper, a system for supporting the creation and management of care plans for children with complex condition. It also presents a preliminary study of GoalKeeper, demonstrating its potential to support families of children with complex conditions. (Chapter 5)
- It develops interactive teaching strategies for agents in a student-teacher reinforcement learning framework, which involve both the student and the teacher in the decision making. It shows that these jointly-initiated teaching strategies reduce the amount of attention required of the teacher agent without compromising the student's learning. (Chapter 6)

Chapter 7 presents conclusions and discussion of future research directions.

Chapter 2

Coordination Challenges in FLECS Teamwork

Complex collaborative activities, such as conducting research, developing software products and treating patients, are often accomplished in a distributed manner, by teams that are loosely-coupled. This chapter describes a qualitative study of teams caring for children with complex conditions. Through this study, we identified a set of teamwork characteristics that make coordination particularly challenging: (1) **F**lat team structure, (2) **L**oose-coupling of activities, (3) **E**xtended duration of the teamwork, (4) **C**ontinued revision of plans, and (5) **S**yncopated time scales of team members. Taken together, these characteristics yield a teamwork setting which we refer to as *"FLECS" teamwork*.

FLECS teamwork differs fundamentally from the teamwork settings addressed by prior work in the social sciences and computer supported cooperative work (*CSCW*). Therefore we analyzed the study finding through the lens of SharedPlans [Grosz and Kraus, 1996], a computational theory of collaboration which provides a specification of teamwork general enough to cover such teams. Based on this analysis, we identified mechanisms lacking in complex care teams and to suggest possible ways technology could provide such mechanisms.

Specifically, we suggest following key roles for technology for supporting FLECS teamwork: (1) making the care plan "ever present" to ensure that team members' individual activities align with the overarching care goals ; (2) supporting efficient information sharing such that team members are not overwhelmed or lack important information, and (3) enabling team members to easily adapt and expand the plan, while ensuring these changes do not lead to conflicts.

The chapter is organized as follows: section 2.1 provides background and related work on care teams; section 2.2 describes the study methodology; section 2.3 describes the study findings; section 2.4 summarizes the study findings and describes an analysis of these findings as well as design implications.

2.1 Background: the Care of Children with Complex Conditions

This section describes the care environment of children with complex conditions and prior work on care coordination.

2.1.1 Care Teams

Children with complex conditions have multiple chronic, simultaneously occurring medical problems. Their care is significantly more complex than care for children with a single chronic disease (e.g., asthma), as it requires expertise from diverse medical specialists and other care providers whose treatments may interact. Care teams for such children are broad in scope, including not only physicians (PCPs and medical specialists) but also other types of healthcare providers (e.g., physical therapists) and people who work with the child in home and community settings (e.g., health aides, teachers). Henceforth, we refer to all caregivers

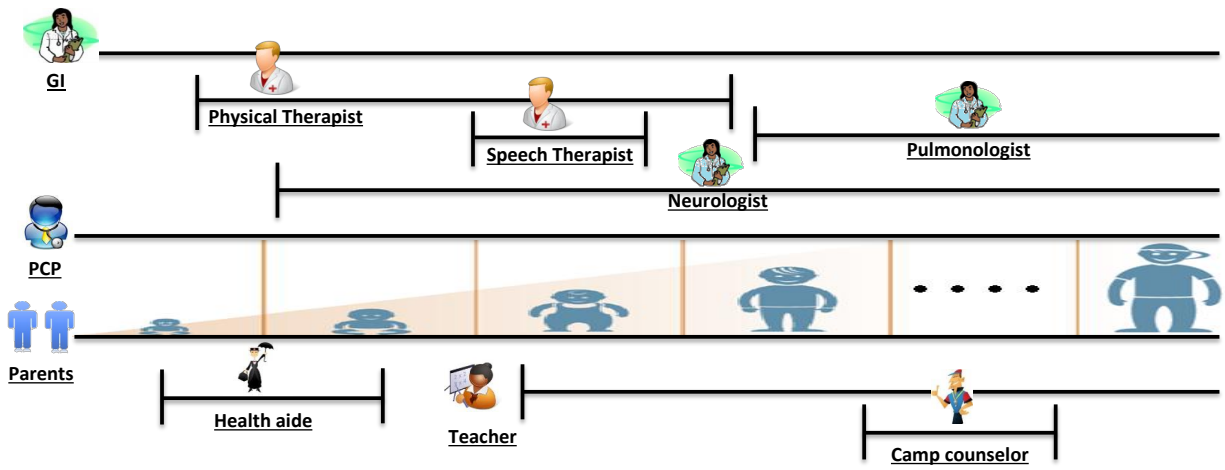

Figure 2.1: The complex care environment

who are not family members as *providers*.

Providers differ in their expertise, and they address different aspects of a child's condition. Their involvement with the child's care may be continuous or intermittent, long or short term, as illustrated by the horizontal lines in Figure 2.1. The group of providers may change significantly over the years, either as a result of personnel changes or because the child's condition or a new developmental stage raise different needs.

Patient-centered care, which has been shown to be critical to improved health outcomes [Sia et al., 2004, Larson and Reid, 2010], requires that patients and families be activated members of the care team. Thus, parents are special members of a child's care team. Although family engagement has been shown to improve health outcomes, this concept is relatively new, and several barriers to engagement stand in the way of its full implementation, including low health literacy and confidence of families as well as the organizational culture of the healthcare system [Carman et al., 2013, Barry and Edgman-Levitan, 2012].

2.1.2 Care Plans

The Chronic Care Model is the dominant blueprint for health-care solutions for adults and children with complex chronic conditions in the US. It argues that effective patient

care requires a collaborative health-care team, inclusive of an engaged, activated patient and family, supported by computer systems and with care framed by an integrated care plan [Coleman et al., 2009]. According to this model, care plans should be organized around whole-person goals (e.g., school readiness) rather than organ-system goals (e.g., brain, lung).

A report by the Lucile Packard Foundation for Children's Health (LPFCH) [McAllister, 2014] provides recommendations for the content of care plans and the process of developing them. It indicates that care plans should include clinical goals (e.g., feeding without a tube), family goals (e.g., going on a family trip) and "negotiated actions" (i.e., actions the team agrees upon) towards achieving the goals (e.g., starting occupational therapy, getting walking support device).

An example of a possible care plan, provided by our clinical collaborators at Stanford, is shown in Figure 2.2. This care plan is for a made-up patient Alex, a 2 month old diagnosed with Down Syndrome. Alex has poor muscle tone and a hole in his heart that make it difficult for him to gain weight and grow. He is currently being fed through a tube down his nose. The care plan specifies three overarching goals for the year: (g_1) move Alex to oral feeds, (g_2) get Alex into daycare, and (g_3) have Alex able to go on a family vacation trip. Figure 2.2 also shows the actions required for achieving these goals and the team members responsible for each. To ensure that progress is made toward the goals and that the plan is revised when needed, the team needs to monitor Alex's status. For example, for goal g_1 they need to track Alex's vomiting and secretions, ability to tolerate home therapy exercises, and ability to take the feeding-related medications.

2.1.3 Inpatient and Outpatient Care Settings

When patients need to be hospitalized due to severe acute conditions, their *inpatient* care is provided by a hospital-based care team. Children with complex conditions receive most of their care in *outpatient* settings. Therefore, our study focused on such settings. The

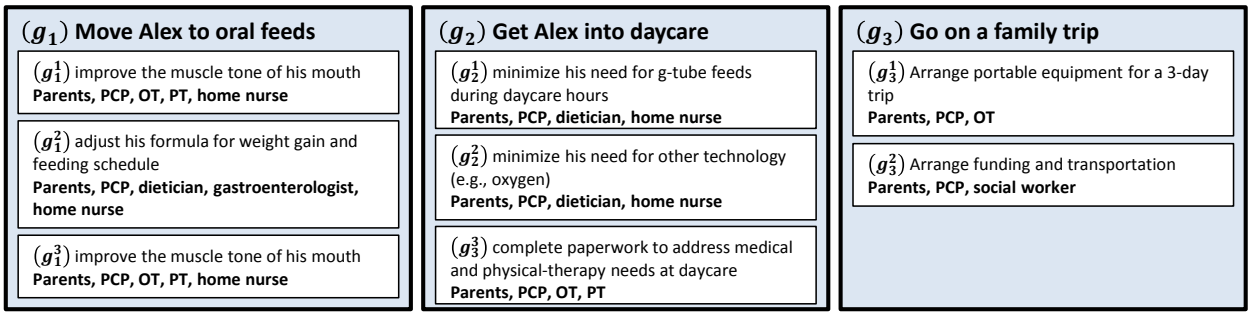

Figure 2.2: Sample Goal-Centered Care Plan.

frequency with which these patients see particular medical providers varies depending on their conditions. The care team is distributed, and team members interact with the child and with each other less frequently than in inpatient settings. It is especially for outpatient periods that care plans need to be effectively deployed and care coordination mechanisms enhanced.

2.1.4 Care Coordination

Coordinated care has been shown to lead to improved health outcomes for children with complex conditions and can reduce healthcare costs [Fiks et al., 2012]. However, coordinating care across care teams is hard and care coordination is often not achieved [Leape, 2012]. Studies of medical practices' use of electronic medical records (EMR) to support care coordination found that while EMRs facilitate improved coordination within a single organization, they were ineffective for coordinating providers in different organizations and were also found to created information overload [O'Malley et al., 2010, Bates, 2010]. According to the study, EMRs support linear processes rather than dynamic coordination processes and are optimized for billing uses rather than for provider and patient needs. In sum, current EMR systems do not provide mechanisms for supporting the coordination of team-based plans.

Human-computer interaction research has investigated systems to support patients in

managing their own care [Klasnja et al., 2010a, Ballegaard et al., 2008], but this work has focused on the patient alone and has not considered other care team members. The CSCW community has studied healthcare teams and developed tools for supporting them [Fitzpatrick and Ellingsen, 2013]. For example, prior work on coordination in inpatient settings has studied temporal and spatial coordination processes in hospital wards [Bardram, 2000, Bardram and Hansen, 2010], trauma-room coordination [Sarcevic et al., 2011], coordination of clinical and non-clinical staff [Abraham and Reddy, 2008], and collaborative information processes in care teams [Reddy and Spence, 2008]. In outpatient settings, prior work has studied coordination in mobile teams (typically of therapists) and developed systems to support the loosely coupled work of such teams [Pinelle and Gutwin, 2006] and systems for supporting therapists' meetings [Kientz et al., 2006]. Our findings go beyond these prior works and show that teamwork in complex care is significantly more complex than in these previously studied settings and thus requires new approaches for supporting coordination.

2.2 Research Settings and Methods

To better understand the challenges of care coordination and the barriers to implementing team-based care plans, we conducted a study over a period of 10 months during 2013–2014 that comprised of observations and semi-structured interviews with parents and providers. Institutional Review Board approval was obtained. Most of the study was conducted in the Complex Primary Care Clinic (*CPCC*) at Stanford's Lucile Packard Children's Hospital, which operates a special complex care program. Interviews with physical and occupational therapists and a social worker were done at a nearby complex care clinic that shares many patients with the CPCC.

Role	N	Data Collection
Parents	13	Individual interviews with 4 parents (children ages ranged between 1.5 to 4 years old); focus group with 9 other parents who are also parent mentors (with children in their teens).
PCP	4	All were interviewed individually. Two of them were also observed for 2-3 hours.
Specialists	4	Individual interviews with a neurologist, pulmonologist, neonatologist and a cardiologist.
Therapists	8	3 focus groups with occupational and physical therapists (2–3 therapists in each interview).
Director of family-centered care	1	Participated in the parent focus group and in a meeting with the complex care program manager.
Care coordinator	1	Observation of 2 hours and informal conversations during that time.
Social Worker	1	Individual interview.
Program directors	2	Meeting with complex care program manager and medical director.

Table 2.1: Study participants.

2.2.1 Participants

In the course of the study, we interviewed and observed representatives of different types of caregivers: parents, parent mentors (who are themselves parents of children with complex conditions), primary care physicians, medical specialists, therapists, a care coordinator, a social worker and administrators. Table 2.1 summarizes the participants by type and data collection methods. All parents we interviewed had children with complex conditions. The care teams for these children included 10–15 providers. For example, the care team for one family included a PCP, gastroenterologist (GI), neurologist, ear nose and throat doctor (ENT), a physical therapist (PT), occupational therapist (OT), speech therapist, and a school-based therapist. Another child's care team comprised 15 care providers including a PCP, cardiologist, liver transplant team, dermatologist, ENT, geneticist, GI, PT, OT, pulmonologist, and rehab specialist.

2.2.2 Data Collection

All interviews, focus groups and observations were conducted during visits to the CPCC in July 2013, November 2013 and March 2014. Notes were taken during interviews, focus groups and observations. The interviews with specialists and social worker and the focus groups with therapists were audio-recorded and transcribed.

Participants of the parent-mentors focus group, which lasted 2 hours, were asked to describe their experiences in managing the care for their children. They were asked to describe challenges that they face and the tools they use to manage their child's care and to brainstorm about tools that would help them better track and manage care.

In a one hour meeting with complex primary care program directors, we asked about care coordination processes and problems and about the types of support they thought would be useful for improving the use of care plans.

Individual interviews with parents lasted about an hour each. Parents were asked about the structure of the care team for their child, the use of care plans and care goals in their child's care, the tools they use to manage and track care, the communication among team members and the challenges they face.

Interviews with PCPs and specialists and focus groups with therapists lasted about 45 minutes each. Interviewees were asked about their patient load, how often they see their complex patients, their use of plans and goals in care, the ways in which they gather information, their communication with parents and with other providers and the challenges they face.

In addition to interviews and observations, we collected relevant documents, including care plan templates, examples of specific care plans and patient information sheets.

2.2.3 Data Analysis

Analyses of interview and observation data (transcripts and notes) were done using affinity diagramming [Beyer and Holtzblatt, 1999]. We iteratively clustered data into themes (e.g. "parents' frustration about lack of communication between providers", "physicians' view of care plans"). We discussed and revised these emerging themes over the course of several sessions.

2.3 Study Findings

This section presents study findings on complex care teams and their use of care goals and care plans. The study revealed the complex FLECS nature of the teamwork in which care teams engage and resulting challenges to implementing long term team-based care plans. Extended duration and loosely coupled caregiver activities are inherent properties of such care. The section thus focuses on the main study findings with respect to other characteristics of FLECS teamwork. It also describes the communication deficiencies that were revealed as communication is key to teamwork.

2.3.1 Consensus-Driven Plan Development

Team-based care plans require that care teams, including the parents and patients (if they are old enough), reach consensus on goals for a child. Our findings indicate, to the contrary, that at present care goals are typically defined separately by individual providers. Furthermore, their use varies among providers: while for some providers setting and tracking goals is a regular part of their practice, others set goals only at times of major health events (e.g., new diagnosis) or do not set goals at all. Conversations between providers and parents about goals vary greatly depending on the family and the provider. Some families have a clear idea of goals whereas others do not feel confident about setting goals. Some providers

discuss goals with parents regularly while others do not discuss goals at all.

Of all providers, occupartional therapists (OT) and physical therapists (PT) are most accustomed to working with care goals. They have a more standardized process for setting and tracking them than other providers. Therapists said that they usually set 2 or 3 goals related to mobility and feeding, which are challenging areas for most infants with complex conditions. A sample PT goal is that of a child achieving the transition from sit to stand. A sample OT goal is that of a child no longer needing a feeding tube. Therapists often define subgoals (i.e., smaller steps) toward achieving a larger goal.

When asked how he works with a family to set goals, one specialist said: "I ask parents about seizures, skin infections. I ask about quality of life – Is the kid sleeping? Is the parent sleeping? Have they developed a strategy? I also ask the parents what is their top priority. I find that single question among the most helpful I ask." For another specialist, discussions about goals usually arise when there is a life shortening diagnosis, near the end of life or when the patient is hospitalized. In these situations she asks the parents "What are [your] goals for [your] child... How much would you put them through for treating this illness." They then decide on a treatment plan depending on the desires of the family.

Many providers discussed the difficulties of setting goals with parents. One therapist said that some parents have very specific goals such as the child not tripping when walking, while other parents have very vague goals such as wanting their child "to be better". A specialist commented, "different patients have a different idea about goals. Many patients are taken aback by it... I tend to give them ideas."

Some of our findings corroborated prior work. For instance, in a study on goal settings for elderly patients [Schulman-Green et al., 2006], providers said that patients sometimes feel uncomfortable discussing goals as they expect physicians to tell them what their goals are. On the other hand, some patients felt that physicians did not have time to discuss goals and did not engage in such high-level discussions.

Our findings also reveal substantial new challenges specific to complex care teams. In particular, the different processes and approaches to goal setting of various providers and the lack of joint goal setting by the team create significant difficulties for parents: They need to prioritize goals because "everyone wants to work on everything", and they also need to track the various goals. While parents frequently discuss goals with therapists, they said that with some doctors "goals don't come up at all" and that they would have liked to discuss the therapists' goals with other providers. These challenges are further exacerbated by the fact that parents have little experience in goal setting: they are often uncomfortable with the idea itself and when they do finally engage, they have difficulty articulating goals that are neither too specific nor too vague.

2.3.2 Continual Distributed Revision of Plans

Our study revealed that current care plans are usually individual provider plans, which are not well integrated into team plans. Parents have difficulties tracking these separate plans. Given the evolving condition of children with complex conditions, plans often become obsolete, and replanning is required. According to both parents and providers, however, even when team plans are implemented, they are rarely consulted or revised.

Furthermore, the nature of care plans varies, with some being simply a list of actions without any clearly specified connection to goals. For example, one PCP's care plan included a list of diagnoses, a high-level assessment of the patient and a list of low-level treatment actions (e.g., take new medication) and required tests (e.g., conducting a sleep study). This list was not organized around goals and was not coordinated with the plans of other team members or developed by the team as a whole. Parents expressed frustration at the use of such plans: "sometimes the care plan is a set of tests and it is not clear what the *plan* is [emphasis ours]."

To implement team-based care plans, "Pediatric Advanced Comprehensive Care Team

(PACT) chats" are being held for patients in the complex primary care clinic. PACT chats include "core members" of the care team as identified by the family. They meet together to discuss care plans for the patient. PACT chats aim to facilitate a setting for the team to create a joint, centralized planning process. As one specialist reported, however, they are "totally not scalable" as they require getting all team members together at the same time, and the meetings themselves take a long time. Therefore, even if the team succeeds in having an initial PACT chat, it is infeasible to have such in person full team meetings repeatedly.

When a severe acute condition arises and a child transitions to inpatient settings (i.e., is hospitalized), a "care conference" might be initiated. In care conferences, a large number of team members meet together in one room to discuss the patient's condition and decide on next treatment steps. Care conferences greatly help everyone "get on the same page", but they occur rarely, usually only when there is a severe acute condition. They mostly involve team members from within the hospital and only some outpatient providers participate.

In outpatient settings, replanning is usually done in a distributed manner. Providers might revise their individual plans when they see a patient, but typically do not coordinate with other providers when they do so. To address this problem, the complex primary care clinic has established "status chats" with a smaller number of team members to follow-up on the patient's care plan. However, as with PACT chats, these meetings are hard to coordinate and do not scale well. One specialist described an additional problem with the ongoing tracking of plans: "One of the issues with the complex care initiative is that the PACT chats and all the status chats have to be provider initiated, and so if you don't remember to do it or there's no one coordinating it, it's like where is it going, where do you even look for it?".

These findings highlight the unique difficulties care teams of children with complex conditions experience in maintaining a coordinated care plan: While team members frequently revise their individual plans for addressing a particular aspect of the child's care, there are rarely opportunities for team members to discuss their plans together. There are currently

no effective alternative processes or tools to support team members in ensuring that their distributed activities are coordinated.

2.3.3 Syncopated Time Scales

Providers differ in their frequency and level of involvement in the plan. PCPs see their patients 3 to 4 times a year, and possibly more during times of acute conditions. Specialists typically see patients 2 to 3 times a year (and when an acute condition occurs). While PCPs are concerned with the overall status of the child, specialists are more focused on monitoring and treating the one facet of a child's condition related to their particular speciality. Therapists meet with patients one to three times a week and are thus more involved than physicians in the delivery of day-to-day care.

As a result of their different timescales and level of involvement in a child's care, providers require different information. Parents reported synthesizing and shaping information differently depending on the provider with whom they are talking. For instance, they might share information related to seizures with the neurologist and information related to feeding with the GI, shaping their choices by what they think is most closely related to that provider's aspect of care. In addition, even when asked similar questions by providers, parents felt that different answers were expected. As one parent said, "a doctor asks if she is walking and expects a yes/no answer; a PT will ask how she is walking and how much progress she has made." Another parent said that the medical team might monitor the child's progress through swallow studies that are typically done only 1 to 3 times a year, while the occupational therapists are much more involved in day-to-day feeding therapies, and that there isn't much conversation between the two.

Providers also differ in the timelines they set for goals and the ways they track progress towards those goals. Therapists set 6–12 month goals, monitor progress and explicitly document goals and progress toward them in their reports of an evaluation that typically

occurs every 6 months. Specialists, on the other hand, said that they usually do not set goal timelines. One specialist said he does not do so because of the uncertainty about the development of the child's condition: "The timeline is the next visit [...] So much of it is beyond their [parents'] control." As for tracking goals, some specialists said they revisit goals at each clinic visit (typically every 4 or 6 months). Another said, "I like seeing my patients quarterly, because then at the sick visits you can deal with just the illness, but when you have a visit where there's nothing going on and they're happy, then you have an opportunity to talk about what the goals are [...] being honest about things."

The different timescales in which providers operate and their different involvement in care pose additional challenges for care coordination. Each of the providers has partial information about the child's status and obtains it at different times. Therefore, it is hard to establish a complete picture of the child's condition at any moment. When providers obtain new information, they need to decide whether that information should be shared with other providers without having complete knowledge of others' plans. In practice, providers reported that such communication is often deficient.

2.3.4 Communication Among the Care Team

We also investigated the ways providers and parents communicate and asked them how they determined the information to communicate, because these processes are essential to effective team-based care plan formation and use.

Medical providers often belong to different organizations and frequently are unable to access each others' medical notes. Hence, they seldom have the full picture of a child's condition. Team members with access to others' records are often overwhelmed by the massive amount of information confronting them. They reported that they often miss important information. One specialist said "I use my own savvy to figure out what happened to the patient since I saw them [...] It's all manual. If something big happened I'll get a

phone call."

Providers described communication with other providers as slow and deficient. Most communication between providers is asynchronous, done through email, letters and notes. Providers who are within the hospital system can copy other providers on notes they enter into the EMR. However, one specialist remarked that "figuring out exactly what [the other provider] want[s] you to read requires that you read [their] whole [individual treatment] plan", and that there is lack of feedback when sending information: "You don't know if [other providers] read it... I get stuff all the time, too, and I don't always review my chart in a timely fashion."

As a result of these communication problems, parents find themselves responsible for transmitting information between providers, and are frustrated by this situation: "We need to relay information back and forth... We wanted [the providers] to be able to talk to each other in one room." One specialist noted a problem that arises when families transmit information between providers: "The family is telling me about what has happened [since the last visit] and they say 'this happened and we went to the doctor', and I really don't know what the doctor thought." Parents are often asked about meetings with other providers but do not remember all of the details from such past meetings. In addition, when new providers join the team or when parents call after-hours, they need to re-explain their child's condition.

Parents also expressed frustration at providers not sharing information with them more proactively. According to parents, communication during transitions from inpatient to outpatient settings is crucial, but often lacking. Parents described transition times as especially stressful and said that they would like to have more information to prepare them and make them more confident about taking care of the child themselves.

Our findings corroborate many prior findings, including the ineffectiveness of EMRs [O'Malley et al., 2010], lack of communication among providers [Stille et al., 2006], and families' frustration with this lack [Strickland et al., 2004]. Our study expands these results, showing

that EMRs are even less effective when providers from multiple organizations are involved, and that identifying relevant information is more difficult for providers in these settings. It also uncovered new problems, for instance, the difficulty of establishing common knowledge when communication is asynchronous and there are no means of determining whether other providers have seen information sent to them. It further identified a crucial communication gap at times of care transitions.

2.4 Discussion

Our study reveals many challenges confronting care teams for children with complex conditions. In this section, we first reflect on these findings in the context of a report on the use of care plans in complex care [McAllister, 2014], showing that many aspirations of the healthcare community are currently not realized. Next, we discuss the unique teamwork challenges for care teams for children with complex conditions that make the implementation of care plans especially hard in this setting. Finally, we discuss implications for design of technology support for care teams.

2.4.1 Care Plans: As They Are, As They Should Be

In support of the drive toward the implementation of integrated, team care plans for children with complex conditions, and to encourage their adoption and use, the Lucile Packard Foundation for children's healthcare (LPFCH) published a comprehensive report entitled "Achieving a Shared Plan of Care with Children and Youth with Special Health Care Needs". In their report, they outlined the "10 Principles for Successful Use of a Shared Plan of Care" [McAllister, 2014]. These principles, listed in the left column of Table 2.2, were identified by a panel of physicians, care coordinators and family advocates experienced in complex care and were informed by prior research on complex care coordination. This section

reflects on our study findings in light of the aspirations for the use of care plans described in the LPFCH report.

Our findings reveal that many of the principles for successful implementation of care plans are currently not met, as summarized in the right column of Table 2.2, and that achieving them will require overcoming several barriers. These barriers cluster into four areas, which we order according to their focus on plans per se.

First, successful care plan implementation requires an integrated care plan with shared goals, implemented as a shared document that is monitored and revised over time (Principles 6, 7 and 9). Our findings reveal, however, that in current practice there are typically multiple individual plans formed by different providers, rather than a single shared team plan, and that providers are typically unaware of each other's plans. Even when team-based care plans are successfully implemented, they are rarely monitored or updated, despite the uncertainty in the outcomes of the initial plan and the inevitable, often unexpected, changes in the child's condition.

Second, successful use of care plans requires that communication among team members be clear, frequent and timely (Principle 2), as does care coordination in general. Our findings highlight many difficulties in achieving the requisite levels of communication and information sharing, including information overload, slow and unreliable communication channels and unclear communication between providers and parents. They show that providers cannot easily access or find all relevant information and thus do not have a full understanding of the child (Principles 3 and 5).

Third, care plans need to address the challenges of parent and family engagement and their integration into provision of care (Principle 1). Prior research on patient-centered care has found that patients and families are often not sufficiently engaged as a result of language barriers, cultural issues, the emotionally overwhelming situation, or providers' lack of experience with engaging parents [Carman et al., 2013, Barry and Edgman-Levitan, 2012].

Our study reveals similar barriers (not reported here as they largely replicate prior findings).

Fourth, care plans need to support transitions in care (Principle 8). In our interviews, both providers and parents described transition situations as very stressful, and families expressed the desire to be given more information and to be better prepared so they would feel more confident about meeting their new responsibilities.

2.4.2 Barriers to Effective Care Plan Implementation

Our study exposed the complex nature of teamwork in complex care, revealing five characteristics that, in combination, distinguish it fundamentally from other teamwork settings. In summary, these are,

- ***F****lat-structure, consensus driven plan development*: goal-setting requires consensus of multiple caregivers, with no single decision-maker "in charge".
- ***L****oosely coupled*: the activities of care providers are largely decoupled, but identifying interactions between their activities is crucial for preventing conflicts.
- ***E****xtended duration*: care plans extend over months to years, during which a child's condition evolves.
- ***C****ontinual distributed revision*: care teams must create and frequently update well-coordinated care plans while rarely (if ever) meeting as a whole.
- ***S****yncopated time scales*: the timescales on which care providers interact with the child vary greatly from several times a week to once or twice a year.

These FLECS teamwork characteristics make the implementation of care plans in these settings particularly challenging. The extended duration of plans combined with uncertainty in outcomes of treatments and changes in the child's condition result in a need for ongoing revision of the goals and plans. The disparate timescales and involvement in the care plan

pose difficulties for providers in keeping track of the child's current condition and plan status, and in coordinating their updated plans with others. Despite their different level of involvement in care, it is crucial that all team members reach consensus about the high-level care goals, as these goals guide their individual treatment plans. It is also important for them to be aware of changes in others' plans that might affect their own plans and to avoid undesired interactions when updating their plans. These challenges are further exacerbated by the fact that, for most providers, participating on the care team for such patients is an exceptional situation, beyond the typical care they were trained to provide — most of their patients do not have multiple conditions nor require interaction and coordination with so many other providers. Some of the providers (e.g., PCPs in complex care clinics) are involved with many different care teams, each with a different team composition, and each operating over an extended period of time, often years.

The challenges faced by the care teams in complex care go beyond those previously studied in healthcare teamwork. In particular, the team coordination challenges they experience are not addressed by existing theories and tools. Prior studies of inpatient team coordination typically investigated the coordination of temporal activities (e.g., scheduling of operations [Bardram, 2000]). In such settings the goals are clear (e.g., find a feasible, preferred schedule) and team members operate in the same timescale (plans are made for a specific day). When replanning is needed it is usually done in a centralized manner by a designated team member (e.g., the charge nurse might reschedule operations if an acute patient arrives [Bardram and Hansen, 2010]) and the changes are communicated by that person.

Prior work on outpatient settings has studied the coordination of mobile home care teams [Pinelle and Gutwin, 2006]. Such mobile teams share some characteristics with complex care teams, namely their activities are typically loosely coupled and they often do not meet together to coordinate. However, these teams are smaller than complex care teams and their

high-level care plan is determined by single person (a case manager). Providers do not need to reach consensus about goals and their cooperation is not strictly required [Pinelle and Gutwin, 2006]. In addition, since they all see patients at their homes, they share a common space where they can leave messages to each other (intentionally or unintentionally).

2.4.3 Foundations for Design of Systems to Support Complex Care Teams

For complex care teams to develop and successfully use long-term care plans, as well as for providers to operate as a team, requires systems able to support FLECS teamwork, and in particular, technology capable of supporting the collaborative, mutual consensus decision-making of a distributed, diverse team whose members operate on different timescales and seldom if ever meet as a full group. These systems must also accommodate team membership changing over the extended duration of complex-care plans.

The social science and CSCW communities have investigated teamwork extensively and developed theories and tools for supporting team coordination. Such work has studied teamwork in various domains, exploring various characteristics of the collaborative setting (e.g., mobile teams [Pinelle and Gutwin, 2006], co-located teams [Reddy and Spence, 2008]), team structure (e.g., hierarchical and non-hierarchical teams [Hinds and McGrath, 2006], team homogeneity) and the nature of tasks performed by teams (e.g., loosely-coupled activities and highly interdependent tasks [Hutchins, 1995]). The FLECS teamwork characteristics exhibited by complex care teams distinguish it from the teamwork studied in prior work. As a result, existing tools and approaches do not fully address the challenges FLECS teams face.

In contrast, research in multi-agent systems (MAS) has developed several models of collaboration that more closely match the characteristics of FLECS teamwork. Each provides a formal specification for the design of computer agents able to robustly act collaboratively

as members of a distributed team; they differ in the facets of teamwork they emphasize. The Joint Intentions (JI) [Levesque et al., 1990] formalization focuses on a specification of the mental attitudes required for teamwork. Planned Team Activity (PTA) [Sonenberg et al., 1992] addresses issues of task allocation and team formation. The SharedPlans (SP) [Grosz and Kraus, 1996, Grosz and Hunsberger, 2006] formalization directly represents partiality and evolution of plans. To analyze the opportunities for technology to support complex care, we chose SP, because it presumes only partial plans, distributed teams acting under uncertainty and the need for plans to evolve dynamically. These assumptions contrast with JI and PTA which assume that the team has a complete, fully expanded plan. SP requires certain group decision-making mechanisms, but not centralized replanning or complete knowledge of all team members' plans. It thus better fits the FLECS teamwork characteristics.

SP has been used to improve the performance of multi-agent computer systems teams [Tambe, 1997], as the model for intentional structure in dialogue systems [Rich et al., 2001], and as the basis of design for collaborative human-computer interface systems [Babaian et al., 2002, Gal et al., 2012]. While designed to guide the development of computer agents, it also provides a framework for identifying coordination mechanisms missing in complex care teams and needed to support their team-based plans.

SP is rooted in the observation that collaborative plans are not simply a collection of individual plans, but rather a tight interleaving of mutual beliefs and (coordinated) intentions of different team members. It specifies the beliefs and intentions required of team members for successful collaborative activities. In particular, SP requires that (SP1) each team member commit to (i.e., form specific intentions regarding) the team's performance of the group activity; (SP2) team members establish agreement on a "recipe" for carrying out the group action and establish mutual belief they are using that recipe; SP allows for recipes to be partial, expanded over time, and revised; (SP3) the group agrees on an allocation of tasks in the recipe according to participants' abilities to carry them out (i.e., decompose tasks and

allocate work appropriately); (SP4) team members commit to performing tasks allocated to them (i.e., adopt intentions to do those tasks); (SP5) team members commit to the success of others in doing their tasks (i.e., adopt intentions that their teammates succeed).

According to SP, only the team members selected for doing a subtask determine and know in detail the recipe for that subtask. For example, the neurologist does not need to know the full details of the PT's plan for getting the child to walk, and the PT does not need to know the full details of the neurologist's plan for treating seizures. SP handles the problem of interaction among loosely coupled tasks through the required commitments to the overall team activity and to the success of teammates (SP1 and SP5) and general axioms of intention. These commitments result in several desired behaviors of team members. In particular, they necessitate communication among team members when any of them comes to believe that plans for subparts of the activity interact or when new information is obtained that could affect others' plans.

Thus, importantly, SP handles the tension between the low communication overhead of loose coupling (and the resultant lack of shared information about plan details) and coordination needs (sharing information that matters because of potential plan interactions) by requiring communication when essential, but not full sharing of all plan details by all participants. The need for such *efficient* communication was evident in our study: some of the providers we interviewed reported that when complete plans or notes are sent to them, they are unable to determine the information most important to consider, and they do not review the information in a timely manner as a result of this information overload.

We illustrate several desired behaviors engendered by SP with examples from complex care settings. The commitments to the overall team activity and to the success of teammates require that team members inform others if they learn that their plans are failing or likely to fail. For example, in the complex care domain, if the physical therapist learns in her session that the child's seizures have worsened, she should notify team members who are working

toward the goal of optimizing seizure medication (in this case the PCP and the neurologist). Similarly, when team members update their plans, they should notify others about changes if their plans might affect others' plans. For example, if the GI decides to start feeding the child by mouth instead of through a tube, she should notify the nutritionist and occupational therapist who also address feeding issues. Further, although not directly involved in feeding, the respiratory therapist should also be notified because breathing and feeding often interact.

While such team behaviors and communication protocols are desired, they are hard to achieve in practice in complex care teams. The examples above assumed that team members had sufficient knowledge about others' plans and about the team's goals to realize that information should be shared. Our findings, however, show that care providers do not have this knowledge. Thus, it is unlikely that team members would exhibit the behaviors SP prescribes. However, SP also suggests ways in which technology could support care teams and help achieve desired team behaviors. The applications of SP in multi-agent systems demonstrate that it is in fact both feasible and efficient to allow individual team members to dynamically modify their individual plans and that only relatively limited communication is necessary to ensure that actions of a team members do not interfere with others' actions.

A SharedPlans-based analysis of complex care team needs suggests the following key roles for technology for supporting complex care teams:

- Support efficient information sharing by team members.
- Make the care plan "ever present", adapting the content and form of its presentation to individual team members based on their involvement in the plan and context of use.
- Enable care team members to easily adapt and expand parts of the plan, while ensuring their changes do not conflict with others' activities.

Efficient information sharing will help team members act in a way that does not conflict with others' plans (SP5). The partiality and dynamic nature that SP assumes, and that

complex care plans exhibit, leads to the need for enabling easy plan adaptation and expansion. Deciding what information to proactively share with team members requires reasoning about the role of that team member in the care plan, and dependencies between different aspects of the plan.

Making the care plan "ever present" will support team members in establishing and maintaining agreement and mutual belief about the high-level team plan and allocation of tasks (SP2 and SP3). Currently, these requirements do not hold in complex care teams as there are no mechanisms to support them. While care providers want their teammates to succeed, they do not have sufficient information about others' plans to act in a way that supports others or at least does not conflict with their activities. The representation of plans also needs to be flexible enough to allow for team members to provide incomplete descriptions of plans that can be easily adapted with time.

2.4.4 Beyond Complex Care Teams

While the study described in this chapter focused on complex care teams, systems able to support FLECS teamwork have the potential to improve the coordination and effectiveness of teams in many other settings within and beyond the healthcare domain. For example, teamwork in rescue and recovery efforts, software development projects, collaborative writing and research collaborations all exhibit some or all of the FLECS characteristics, and would benefit from greater support of team coordination [Haake and Wilson, 1992, Yamauchi et al., 2000].

	Principle	**Study Findings**
1	Children, youth and families are actively engaged in their care.	Engaging families is a complex process: families are overwhelmed by the child's condition; family engagement is a new concept and many providers do not feel comfortable with it.
2	Communication with and among their medical home team is clear, frequent and timely.	Communication is typically infrequent and slow. Providers often do not communicate among themselves, leading to lost information and to parents being responsible for transmitting information between providers.
3	Providers/team members base their patient and family assessments on a full understanding of child, youth and family needs, strengths, history, and preferences.	Team members are often focused on their aspect of the care and do not have a shared "big picture" view of the child's overarching goals.
4	Youth, families, health care providers, and their community partners have strong relationships characterized by mutual trust and respect.	Our study did not focus on this issue, but our findings show team members do not all know each other.
5	Family-centered care teams can access the information they need to make shared, informed decisions.	Team members cannot access all of the information about the patient. When they can access it, they are typically overwhelmed by the amount of information and have difficulties finding the *relevant* information.
6	Family-centered care teams use a selected plan of care characterized by shared goals and negotiated actions; all partners understand the care planning process, their individual responsibilities, and related accountabilities.	Different team members have different concepts and uses of care plans and care goals. There is no clear process for defining and revising team care plans.
7	The team monitors progress against goals, provides feedback and adjusts the plan of care on an ongoing basis to ensure that it is effectively implemented.	Progress towards goals is often not continuously monitored and care plans are not updated frequently. In addition, different providers set goals separately rather than as a team. Thus, even when monitoring is done it does not contribute much to achieving coordinated care.
8	Team members anticipate, prepare and plan for all transitions (e.g., early intervention to school; hospital to home; pediatric to adult care).	Parents feel they lack information in times of transitions, which they find especially stressful as responsibility shifts to them (e.g., in- to outpatient care) or when new providers join the team.
9	The plan of care is systematized as a common, shared document; it is used consistently by every provider within an organization and by acknowledged providers across organizations.	Different team members have different concepts and uses of care plans and care goals. Typically there are many different plans rather than a single team plan.
10	Care is subsequently well coordinated across all involved organizations/systems.	Care coordination is very hard to achieve, both between organizations and within organizations.

Table 2.2: Principles for successful use of care plans (left column) and findings from our study that point out the barriers in achieving them (right column).

Chapter 3

Mutual Influence Potential Networks: Reasoning About Information Sharing

Of the FLECS teamwork characteristics, the *loose-coupling* and *extended-duration* of the teamwork in particular make effective information sharing hard. By decomposing the group activity into tasks carried out by individual team members, loosely-coupled teamwork reduces the need for negotiation and resolution of conflicts [Olson and Teasley, 1996]. While such decomposition allows collaborators to focus on their individual tasks, it makes identifying dependencies and conflicts harder [Hutchins, 1995, Grosz and Kraus, 1996]. The extended-duration of the teamwork further exacerbates the problem, as plans and dependencies between tasks may change. As shown by our study of complex care teams and by others [Hutchins, 1995], team members in such settings often either lack information about relevant activities of others or are overwhelmed by the amount of information available and unable to identify the subset of information that is important to them. The overwhelming amount of information and the lack of information can both result in coordination failures.

This chapter presents a formalization of a new multi-agent systems problem, Information Sharing in Loosely-Coupled Extended-Duration Teamwork (ISLET). It presents a new

representation, Mutual Influence Potential Networks (MIP-Nets) and an algorithm, MIP-DOI, that uses this representation to determine the information that is most relevant to each team member. Importantly, because the extended duration of the teamwork precludes team-members from developing complete plans in advance, the MIP-Nets approach, unlike prior work on information sharing, does not rely on a priori knowledge of a team's possible plans. Instead, it models collaboration patterns and dependencies among people and their activities based on team-members' interactions. Last, it presents the results of an empirical evaluation in an environment that simulates collaborative activities, showing that MIP-DOI is able to identify relevant information. It further explores the effect of different aspects of the teamwork, including team size and coupling of activities on the performance of the algorithm.

3.1 The ISLET Problem

An *ISLET problem setting* comprises the following:

- P: a set of collaborating partners. The set can change over time with partners joining or leaving the team.
- O: a set of objects that partners interact with. The set can change over time as a result of partners' actions.
- A: the set of act-types $\{ADD, MOD, DEL\}$ for adding, modifying or deleting objects. These general domain independent act-types are specialized to domain-specific act-types in each application domain.
- S: interaction sessions of partners. A session $s(p, t, (\langle a_1, o_1\rangle, ..., \langle a_{|s|}, o_{|s|}\rangle))$ is defined by a triple: the partner acting, the time of the session, and a set of pairs of act-types

and the objects they operate on ($\langle a_i, o_i \rangle$) [1]. For brevity, we denote a session recorded at time t as s_t.

The *ISLET problem* is to determine a set of objects $O_{share} \subset O$, where $|O_{share}| = l$, to inform $p \in P$ about, given sessions s_1 to s_{t-1} and the identity of the partner p who is starting s_t. The constraint on the cardinality of O_{share} (l), is a communication budget, which restricts the amount of information that can be shared. It reflects the need not to overwhelm partners with too much information. The objects in the set O_{share} should be *relevant* to the partner. The notion of relevance has been widely discussed in the literature on cognition and communication [Sperber and Wilson, 1987]. Intuitively, information is relevant if it will affect the partner's actions. The specific definition of relevance, however, is domain dependent.

To illustrate ISLET settings, we will use the example of a collaborative writing scenario in which a group of researchers (the P), comprising Alice, Bob and Chris, writes a grant proposal together. In this scenario, the set of objects (O) includes the paragraphs of the proposal. Specializing to the domain and applying act-types (A) to objects yields such actions as writing new paragraphs, removing paragraphs or editing paragraphs. Sessions (S) are added over time as Alice, Bob, and Chris edit the document, and the set O evolves as paragraphs are added or deleted. P can also evolve over time; for instance, Dan might join in writing the proposal.

On Monday morning (t_{10}), Alice (p_1) edits the document, taking the following actions: modifying paragraph 3 ($\langle MOD, o_3 \rangle$), deleting paragraph 4 ($\langle DEL, o_4 \rangle$) and adding a new paragraph ($\langle Add, o_5 \rangle$). These actions together constitute the session shown in Figure 3.1(a). The following day, Chris begins editing the document. In this example, the ISLET problem is to choose the set of paragraphs to share with Chris. For example, if $l = 2$, O_{share} should include the two paragraphs that have changed since Chris last edited the document and that

[1] We use only the $\langle a_i, o_i \rangle$ pairs to emphasize that the partner and time are the same for all actions taken in a single session.

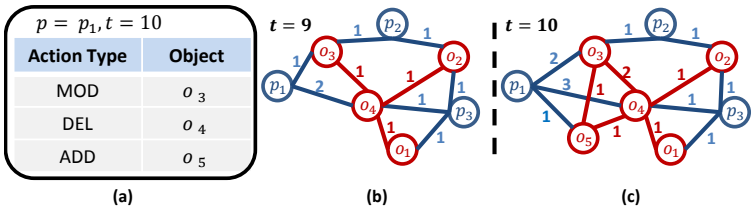

Figure 3.1: (a) An interaction session s_{10}; (b) The MIP-Net after sessions $s_1 - s_9$; (c) The updated MIP-Net after session s_{10}.

are deemed most relevant to Chris' activities.

3.2 Mutual Influence Potential Networks

MIP-Nets represent interactions between partners and objects and dependencies between different objects. In a MIP-Net, partners and objects are represented by nodes. A particular partner $p \in P$ and an object $o \in O$ are represented by nodes n_p and n_o, respectively. Henceforth, we use **p** when referring to a particular partner and **o** for a particular object.

Particular nodes n_p and n_o are connected by an edge if **p** performed an action on **o**. The edge weight corresponds to the extent of the interaction: if **p** takes many actions that affect object **o**, this will be reflected by a high weight on the edge connecting n_p and n_o. Thus, the weights on such edges represent information about team members' responsibilities, which we refer to as "role allocation". Similarly, n_o and $n_{o'}$ are connected by weighted edges based on the frequency at which the objects they represent are modified in the same sessions. Edges connecting object nodes thus represent object dependencies, i.e., the extent to which team members tend to change one object when they change the other. We refer to these object dependencies as the "task structure", because these groupings are likely to be a reflection of an underlying task. For instance, in a research paper, paragraphs reporting results in a Results section and in a Conclusion section might be frequently edited together as part of the same underlying task of adding new results to the paper. Importantly, this sense of task structure is much looser than that used in formal plan representations such as Hierarchical

Task Networks in which dependencies between different tasks are explicitly specified.

Formally, a MIP-Net consists of:

- N_P: a set of partner nodes.
- N_O: a set of object nodes.
- E: a set of edges, each edge connecting a partner node with an object node or two object nodes.

Figure 3.1(b) shows a sample MIP-Net. Partner nodes and edges connecting partners and objects are shown in blue. Object nodes and edges connecting them are shown in red. Numbers on edges represent the edge weights.

3.2.1 Constructing and Updating MIP Networks.

MIP-Nets are constructed and updated over time based on partners' sessions. At the end of each session s_t, the MIP-Net is updated. The MIP-Net update procedure, shown in Algorithm 1, first checks whether $\mathbf{p}$ is already represented by a node in the MIP-Net. If not, a new node is added to N_P (lines 1–2). Next, it iterates over all actions in the session; new object nodes are added as a result of ADD actions, and the weights of edges connecting n_p with object nodes representing objects on which that partner acted are incremented by 1 (lines 4–8). Similarly, the weights of edges connecting object nodes representing objects that the partner interacted with in the same session are incremented by 1 (lines 9–15). We note that nodes representing deleted objects persist in the MIP-Net as information about their connections can implicitly reveal dependencies between other objects.

To illustrate the MIP update procedure, consider the collaborative writing scenario described earlier: assume the MIP-Net at time $t = 9$ is the one shown in Figure 3.1(b). Following s_{10} (Figure 3.1(a)), the MIP-Net is updated, yielding the network shown in Figure 3.1(c). As shown, a node representing o_5 was added to the MIP-Net and the weight on

Input: $s(p, t, (\langle a_1, o_1 \rangle, ..., \langle a_{|s|}, o_{|s|} \rangle))$

if $n_p \notin N_P$ **then**
  $N_P = N_P \cup n_p$
end
for $a, o \in s$ **do** `// increment weights of p-o edges`
  if $a = ADD$ **then**
    $N_O = N_O \cup n_o$
  $IncrementWeight(n_p, n_o)$
end
for $a, o \in s$ **do** `// increment weights of o-o edges`
  for $a', o' \in s$ **do**
    if $o \neq o'$ **then**
      $IncrementWeight(n_o, n_{o'})$
    end
  end
end

Algorithm 1: The MIP-Net update procedure.

edges connecting p_1 (the node representing Alice) with o_3, o_4 and o_5 were incremented. The weights on edges connecting all pairs of objects included in the session (e.g., o_3 and o_4) were also incremented. Although o_4 was deleted, the node representing it persists in the MIP-Net.

The computational complexity of this procedure is dominated by $|s|^2$, where $|s|$ is the number of $\langle a, o \rangle$ pairs in the session. The update procedure requires one iteration over the set of $\langle a, o \rangle$ pairs to update the weights connecting n_p with nodes representing the objects interacted with during the session, and a second iteration over all pairs of objects o, o' that were interacted with in the session to update weights on edges connecting object nodes.

3.3 The MIP-DOI Algorithm

The MIP-DOI algorithm uses the MIP-Net to reason about information sharing in ISLET problem settings. To quantify the relevance of modifications to some object **o** to some partner **p**, we use the concept of *Degree-Of-Interest* (DOI). Furnas [1986] defined $DOI(x \mid y)$ as the degree of interest a user has in an item x, given that the user is focused on some item y. It is

computed it as follows:

$$DOI(x \mid y) = \alpha \cdot API(x) + \beta \cdot D(x, y)$$

$API(x)$ is the *a priori* importance of item x. It is independent of the user's identity and aims to reflect the *global* importance of an item. $D(x, y)$ is the distance between x and y. It aims to reflect the importance of x given the user's context.

The rationale behind this formulation is that, generally, a user will be interested in items that are close to her current focus of attention, as well as in items that are of general importance [Furnas, 1986]. This notion of DOI fits our purposes, as collaborators will likely find value in information about objects that are closely related to objects they interacted with or currently focus on, as well as in information about objects that appear to be of significant importance to the team's activities as a whole.

As initially introduced, DOI was computed over items in a tree. Similar to Van Ham and Perer [2009], we use a network-based DOI metric. In our formulation of DOI, we consider two different nodes as representing **p**'s focus of attention: (1) the node representing the partner in the MIP-Net (n_p), as the edges from n_p capture the extent of interaction between **p** and the different objects, and (2) the node representing the object that the partner acts on at the beginning of a session, denoted o_f for "focus object". In many settings, information about o_f is available to the system (e.g., observing the paragraph Alice starts editing) and can be integrated in the DOI computation. In sum, we measure DOI by computing:

$$DOI(o \mid p, o_f) = \alpha \cdot API(n_o) + \beta_1 \cdot D(n_o, n_p) + \beta_2 \cdot D(n_o, n_{o_f})$$

The distance values $D(n_o, n_p)$ and $D(n_o, n_{o_f})$ can be computed using various distance measures for networks. We used the Adamic/Adar proximity metric [Adamic and Adar, 2003], adapted to take into account edge weights.

Network centrality metrics can be used to compute the *a priori* importance of an object node n_o. Our implementation uses $deg(n_o)$ (the sum of weights on edges connected to n_o). Note that the importance of objects can change over time. For instance, if many partners interact with an object, its degree will increase and thus its centrality will increase.

To determine the set of objects $O_{share} \subset O$ to share with $\mathbf{p}$, the MIP-DOI algorithm computes $DOI(o \mid p, o_f)$ for each $o \in O$ and chooses the l objects with the highest DOI (recall, l is the communication budget). The computational complexity of MIP-DOI depends on the methods used to compute API and D. In our implementation it is dominated by $|O|^2$.

3.4 Empirical Methodology

Before integrating MIP-DOI intro real-world systems (as described in Chapter 4), we evaluated the ability of MIP-DOI to identify relevant information with a collaborative activity simulation. This simulation was conducted for two reasons. First, it provides a ground truth for assessing the relevance of information. Second, collaborative activities can vary in many aspects, including the size of the group, frequency of interactions and coupling of tasks. For example, Wikipedia articles are written by a large number of authors with a small percentage of the authors making the majority of contributions, while academic papers are typically written by a much smaller number of authors who act in a more coordinated way (e.g., they might divide responsibilities for different sections). Software projects hosted in GitHub also differ significantly in the nature of the collaboration on projects [Kalliamvakou et al., 2014]. Some projects include a small group of collaborators that contribute fairly equally, while others have one or two main contributors and a large number of developers who only make a single contribution. In healthcare, the role allocation among care providers is much more strict due to their specialization. The simulation enables exploration of the effects of such aspects of teamwork in a controlled environment.

3.4.1 Collaborative Activity Simulation

We designed a collaborative activity simulation in which a group of partners (P) are faced with a constraint satisfaction problem that abstracts the type of coordination problems that arise in collaborative activities. In the simulation, the partners collaboratively color a graph $G(V, E)$ using a set C of colors such that no two neighboring vertices are assigned the same color. Constraints on the colors of neighboring vertices correspond to a group's need to align their activities. For example, in the writing scenario, a paragraph summarizing the results in the introduction of the paper must align with the results described in the results section. In healthcare, a choice of a course of treatment for one condition can constrain treatment of other conditions due to conflicting effects.

We formulate this collaborative activity as an instance of an ISLET problem as follows:

- P: collaborating partners.
- O: graph vertices.
- A: The act-types MOD, DEL and ADD are instantiated as follows: $mod(v, c, c')$ changes the color of v from c to c', where $c, c' \in C$. $add(v)$ adds a new vertex v' as a neighbor to an existing vertex v. $del(v)$ removes vertex v from the graph.
- S: Interaction sessions: the session $s(p, t, (\langle a_1, o_1\rangle, ..., \langle a_k, o_k\rangle))$ consists of the changes made to the graph by p at time t.

For simplicity, in this section we describe a simulation in which the set of objects is constant (i.e., only the MOD act-type is used).

To reflect the information that would be available in real-world settings, the algorithms do not have access to the graph structure (G). They only know about the existence of objects (vertices) that partners interacted with and their colors, but do not have information about edges.

Partners know the graph structure (i.e., the edges between vertices), but do not know the current color of a vertex unless it was shared with them, and they assume a vertex's color has not changed until they receive new information. This reflects the ISLET setting in that partners might know what potential dependencies exist between different objects, but not the current state of the different objects. They thus might not be aware of conflicts. For example, Alice might know that there is mutual dependency between different sections of a proposal, but not be aware of inconsistencies in current versions of the sections without reading them.

Two key factors in the simulation are the graph structure and the way partners choose O_{modify} at each round. As described earlier, we aim to support loosely-coupled teamwork. In such teamwork, we expect partners to have some (possibly non-strict) allocation of roles, and that there would typically be more constraints between objects that are modified as part of a specific role or task, than between objects that are modified in the context of different roles or tasks.

To model the loosely-coupled nature of the teamwork, we generate clustered graphs, where vertices within each cluster are more likely to have an edge connecting them than vertices that belong to different clusters. We specify the number of clusters, the probability of creating an edge between vertices in a cluster (p_{within}) and the probability of creating an edge between vertices in different clusters ($p_{between}$). We then randomly generate graph instances using these parameters. For example, there are likely to be more dependencies between paragraph in the same section than between paragraphs in different sections. Similarly, in complex health care, there are likely to be more dependencies between treatments related to one aspect (e.g. mobility) of the care, than between treatment related to different organ systems.

To reflect role allocation among the partners, we assign each partner with probability distribution over vertices. Each partner is assigned a single primary cluster, and most of the probability mass ($p_{primary}$) is assigned to that cluster. The choice of the *focus object* (o_f) is based on these distributions: with probability $p_{primary}$, **p** will choose an object o_f from its

primary cluster, while it will choose o_f from another cluster with probability $1 - p_{primary}$. The remaining $k-1$ vertices in O_{modify} are chosen in proportion to their distance from o_f in the graph. Specifically, the probability of choosing a vertex decreases exponentially with its distance from the o_f. The rationale for this choice of objects is that partners are more likely to choose objects related to their primary task, and further that they are more likely to modify a set of objects that interact with each other. For example, when editing a document Alice will have a task in minde and will not simply choose a random set of 5 paragraphs to edit, but rather edit paragraphs that are related to some higher-level aspect of the paper).

In each round of the simulation procedure, shown in Algorithm 2, the partners take turns modifying vertex colors, as follows: (1) In turn, each partner $\mathbf{p}$ chooses a *focus object*, denoted o_f, and a set of k objects to modify denoted O_{modify} (line 3). The object o_f is chosen from the partner's primary cluster with probability $pr_{primary}$ (and from a different cluster with probability $1 - pr_{primary}$). The remaining $k-1$ vertices in O_{modify} are chosen in proportion to their distance from o_f in the graph, to reflect higher likelihood of partners carrying out activities that are closely related to each other in each session; (2) A set O_{share} of l objects to inform $\mathbf{p}$ about are chosen by the information sharing algorithm, given $\mathbf{p}$, o_f and sessions s_1 to s_{t-1} (line 4); (3) The belief of $\mathbf{p}$ about vertices' colors is updated to reflect the shared information (line 5); (4) $\mathbf{p}$ chooses colors for objects in O_{modify}, such that the assignment minimizes the number of conflicts known to $\mathbf{p}$, based on its updated belief (line 6); (5) The problem instance is updated to reflect the new coloring (line 7).

Evaluation Metrics

We consider an object $\mathbf{o} \in O_{share}$ relevant if there is an edge connecting $\mathbf{o}$ to at least one object in O_{modify}, as such information can directly affect $\mathbf{p}$'s choice of action. We measure precision ($\frac{|O_{relevant} \cap O_{share}|}{|O_{share}|}$) and recall ($\frac{|O_{relevant} \cap O_{share}|}{|O_{relevant}|}$).

Input: $P, problemInstance, k, l, maxRounds$

while $t < maxRounds$ **do**
  for $p \in P$ **do**
    $O_{modify}, o_f = p.chooseObjects(k)$
    $O_{share} = \quad getObjectsToShare(l, o_f)$
    $p.updateBelief(O_{share})$
    $s_t = p.chooseActions(O_{modify})$
    $problemInstance.update(s_t)$
  end
  $t = t + 1$
end

Algorithm 2: The graph coloring simulation procedure.

Parameter	Description
$\|P\|$	Number of partners [5]
$\|Cl\|$	Mean cluster size [10]
$pr_{primary}$	Probability o_f is chosen from the primary cluster [0.8]
pr_{within}	Probability of creating an edge between vertices in the same cluster [0.3]
$pr_{between}$	Probability of creating an edge between vertices in different clusters [0.05]
$k = \|O_{modify}\|$	Number of actions p can take in a single session [3]
$l = \|O_{share}\|$	Number of objects that can be shared in a single session

Table 3.1: The parameters controlling simulation configurations. Values in brackets were used in the experiments described in Section 10.

Algorithm Comparisons

We evaluate the performance of the following algorithms:

- **Omniscient**: has access to the graph structure and chooses objects in proportion to their distance from o_f.
- **Most frequently changed**: chooses objects that were changed most *frequently* by partners.
- **Most recently changed**: chooses objects that were changed most *recently* by partners .
- **Random**: chooses the objects to share randomly.
- **MIP-DOI**: varying the coefficients α, β_1 and β_2. We focus on the following configurations to test the effect of the different DOI components, and describe a few combinations of the components:
  - **MIP-DOI-centrality**: the DOI computation only considers objects' centrality ($\alpha = 1$).
  - **MIP-DOI-partner**: the DOI computation only considers objects' proximity to the partner node ($\beta_1 = 1$).
  - **MIP-DOI-focus**: the DOI computation only considers objects' proximity to the focus object node ($\beta_2 = 1$).

With all MIP-DOI configurations, we used Algorithm 1 to update the MIP-Net at the end of each session. To ensure a fair comparison, all algorithms (MIP-DOI and baselines) choose the l vertices to share with $\mathbf{p}$ from the set of objects that were changed last by some $p' \neq \mathbf{p}$. We use the same seed when generating random numbers for determining stochastic decisions (e.g., the choice of o_f) such that all algorithms are evaluated using the same conditions.

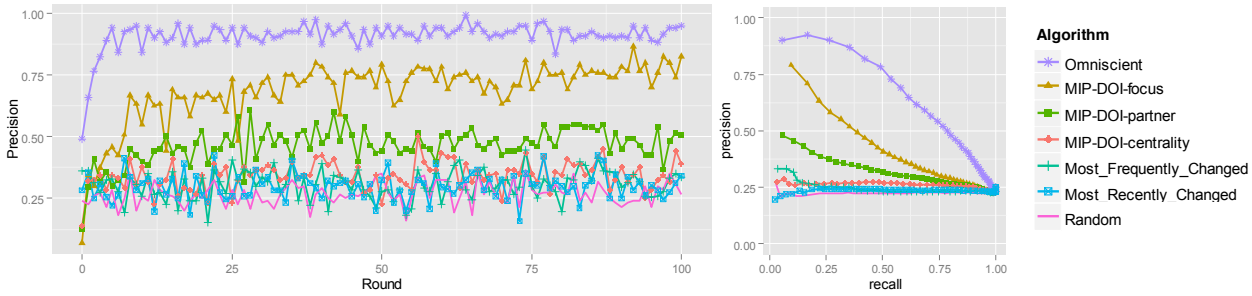

Figure 3.2: (a) Average precision by round (10 different graph instances with 5 runs each). (b) Precision-recall curve generated by varying l; each point shows the precision and recall for a given communication budget (l) with results aggregated from rounds $t_{15} - t_{99}$.

Simulation Results

This section reports in detail the results of a simulation that used the parameter values shown in brackets in Table 3.1. The relative performance of the different algorithms was consistent across other parameter settings. Figure 3.2(a) shows the precision obtained by each of the algorithms with $l = 3$. Overall, all MIP-DOI configurations significantly outperformed all baselines except of course for the omniscient baseline which has access to the graph structure. As can be seen in the figure, of the MIP-DOI configurations, MIP-DOI-focus achieved the best performance. Over time, its performance becomes close to that of the omniscient algorithm as more information about the task structure is accumulated in the MIP-Net.

If algorithms do not have access to o_f, MIP-DOI-partner (proximity of objects to partners) still outperforms all other uninformed baselines, demonstrating that MIP-Nets effectively recover information about partners' role allocation (i.e., their cluster assignment). MIP-DOI-centrality, despite not incorporating the proximity of objects to o_f or $\mathbf{p}$, still outperforms the other baselines, but achieves relatively low accuracy.

Figure 3.2(b) shows precision-recall curves for the algorithms. The curves were generated by varying the communication budget l between 1 (the leftmost points in Figure 3.2(b)) and

	$\beta_2 = \mathbf{1}$ (focus)	$\alpha = \mathbf{0.3}, \beta_2 = \mathbf{0.7}$	$\beta_1 = \mathbf{1}$ (partner)	$\alpha = \mathbf{0.3}, \beta_1 = \mathbf{0.7}$
$\mathbf{t_0 - t_{14}}$	0.40	0.45	0.33	0.36
$\mathbf{t_{15} - t_{99}}$	0.69	0.55	0.42	0.39

Table 3.2: Average precision obtained by MIP-DOI with different configurations in early and late rounds of the simulation.

the total number of changed objects considered for sharing. The results are aggregated starting from round 15, a point at which the MIP-Net has accumulated some information about partners' activities. As can be seen in the figure, all configurations of MIP-DOI significantly outperform the uninformed baselines. The gap between the performance of MIP-DOI-focus and the omniscient algorithm is relatively small when using very limited communication budgets ($l \leq 3$), demonstrating that the MIP-Net representation can effectively distinguish between clearly relevant objects (high proximity to o_f) and clearly irrelevant objects (low proximity to o_f). For larger values of l, the MIP-Net representation is less capable of separating relevant and irrelevant objects and the difference between MIP-DOI and the omniscient algorithm is greater.

While MIP-DOI-centrality does not perform well, integrating α (object centrality) with either β_2 (proximity to o_f, when o_f is known) or with β_1 (proximity to the partner, when o_f is unknown) leads to improved performance in early rounds, as objects that are more central are likely to have more short paths connecting them with other objects, and thus higher probability of being chosen for O_{modify}. This can be seen in the first row of Table 3.2. However, once sufficient information about specific objects and partners' roles is accumulated, integrating α results in lower precision (second row of Table 3.2).

Team members are likely to have more difficulty identifying relevant information about objects they interact with infrequently. Therefore, we examined the extent to which MIP-DOI is able to retrieve relevant objects that do not belong to partners' primary clusters. When using MIP-DOI-focus with $l = 3$, 72% of the objects in O_{share} were from outside of the partners' primary clusters. Using MIP-DOI-partner leads to less sharing of information from

outside the primary cluster (50%), as the DOI focuses on distance from the partner's node. MIP-DOI-centrality shares the most information from outside the primary cluster (87%), but at the cost of sharing many irrelevant objects.

These analyses were based on a specific configuration of the simulation, but the general trends in performance were robust across different parameter configurations of the simulation. We next describe the effects of varying the simulation parameters on the performance of MIP-DOI.

Team size: varying the number of partners ($|P|$) does not substantially affect the performance of MIP-DOI-focus. More objects are changed at each turn, resulting in higher precision when using MIP-DOI-focus. Recall, however, does not increase because there are overall more relevant objects. The performance of MIP-DOI-partner degrades with increased team size, as it takes longer to learn the role allocation.

Cluster size: increasing the number of objects in each cluster ($|Cl|$) leads to lower precision of all MIP-DOI configurations, as it takes the MIP-Net longer to capture the dependencies (constraints) between objects and the roles of partners.

Number of modified objects: when increasing the number of objects a partner can change in a session (k), there are two effects: on the one hand, more information is incorporated in the MIP Update procedure (as more actions are taken). On the other hand, the relationship between pairs of objects is less indicative of constraints between them (e.g., there is a higher likelihood of choosing more distant objects to change together with o_f). Overall, the performance of MIP-DOI is similar across different values of k. Precision increases as there are simply more relevant objects, but recall does not.

Role allocation strictness: the strictness of role allocation is determined by $pr_{primary}$, that is, the probability that a partner chooses o_f from its primary cluster. The performance of MIP-DOI-partner is affected most by the changes to role allocation: with more strict role allocation (higher $pr_{primary}$), it is easier to capture the roles of different partners, and thus

the proximity between object nodes and the partner node is more indicative of relevance. The other algorithms are not affected much by these changes. Their precision slightly decreases when increasing $pr_{primary}$ as less relevant objects change between each partner's consecutive sessions, but recall remains similar.

Graph structure: the parameters pr_{within} and $pr_{between}$ determine the likelihood of edges (constraints) connecting vertices in the same and in different clusters respectively. Generally, increasing both probabilities means that there are more edges in the graph, and thus more potentially relevant objects to share. Therefore, precision generally goes up with higher values of pr_{within} and $pr_{between}$, while recall does not. The exact effect depends on the specific values of these probabilities. $pr_{between}$ in essence controls the level of coupling between partners' activities. With smaller values of $pr_{between}$, it becomes harder for MIP-DOI to learn about the dependencies between different partners' activities, and thus it becomes harder to share with a partner relevant information from *outside* that partner's primary cluster.

3.5 Related Work

Prior work has developed various algorithms for determining information sharing in settings with multiple *agents*. We review key approaches from the multi-agent systems literature and distinguish the problem setting and the approach presented in this chapter from those studied in prior work.

Theories of teamwork and collaboration [Grosz and Kraus, 1996, Cohen and Levesque, 1990, Sonenberg et al., 1992] emphasize the key role of communication in teamwork. Belief-Desire-Intention (BDI) approaches to multi-agent planning often base their communication mechanisms on these theories. For example, the joint intentions model [Cohen and Levesque, 1990] defines conditions for communications, such as communicating to establish joint intentions, communicating the achievement of a goal, and communicating when learning that a goal

cannot be achieved. These ideas were used in the STEAM multi-agent framework [Tambe, 1997]: In this framework agents use a decision tree to determine whether to communicate information about operation termination by considering their belief about the joint intentions of the team and weighing (the domain specified) costs and risks of not communicating. Other works have developed teamwork programing languages that include explicit rule-based communication mechanisms [Weerasooriya et al., 1995, Pokahr et al., 2005].

Prior work on decision theoretic approaches to multi-agent communication can generally be classified to two types: approaches that reason about communication during planning time, and approaches that reason about communication during execution time. The DEC-POMDP-COM model [Goldman and Zilberstein, 2003] and the COM-MTDP model [Pynadath and Tambe, 2002] provide a theoretical model for reasoning about communication during planning time and include communication actions in the agents' policies. Spaan et al. [2006] developed a model in which the communicated information is included in the actions vectors of agents, and is then incorporated into the observation vectors received by agents in the next time step. Offline approaches assume that all possible observations that agents may receive are known during planning time.

Other approaches reason about communication during execution time [Xuan et al., 2001, Oliehoek et al., 2007, Emery-Montemerlo et al., 2005]. Such approaches aim to identify situations in which communication would improve the group's performance, based on observations obtained by agents. This reduces computational complexity since agents do not need to consider in advance what to communicate at each possible scenario, but rather only reason about communication given their actual observations. For example, Roth et al. [2005] reason about communication by growing a tree of the possible joint beliefs of the team. This work has been extended to consider not only when to communicate, but also what subset of observations to communicate to other team members [Roth et al., 2006]. The approaches proposed in the above two works assume a known joint policy, such

that agents can deterministically determine what action each agent in the team would take following communication. Wu et al. [2009] propose an algorithm for online planning and communication. The algorithm merges observation histories based on their similarity in terms of future actions. They then use these history cluster to reason about histories incompatibility, which determines communication. This model does not include communication cost, but instead tries to maximize utility while minimizing the amount of communication. Several other approaches have also addressed the problem of communication during execution time.

There are few prior works that have combined BDI concepts with decision theoretic approaches to reason about communication. Inspired by BDI teamwork, Kwak et al. [2011] define trigger points in which communication should be considered in the context of a DEC-POMDP model. These trigger points occur when there is an ambiguity in the mapping from a joint policy of an agent to its action. When such ambiguities arise, agents reason about the possible gains from communication, where an agent can communicate either by *asking* for information or by *telling* other agents about their observations. This model still requires knowledge of all possible observations that might be obtained during execution. Finally, Kamar et al. [2009] developed the PRT representation of collaborative activities which extends the BDI SharedPlans theory to include uncertainty and utility. They further develop algorithms that evaluate the utility of PRTs to reason about helpful actions, including communication. These algorithms compare the utility of a given PRT representing probabilities over possible plans of collaborating agent, with the utility of a modified PRT that takes into consideration the new information and might therefore assign different probabilities to plans.

All of the approaches described above, both BDI and decision theoretic, rely on a *complete plan knowledge assumption*; they assume availability of a complete domain model of the actions or plan library, state space, and utilities or goals. They use this model and knowledge of a team's plans or policies to compute the value of information. Although some approaches

assume only incomplete knowledge of agents' plans and use reinforcement learning [Zhang and Lesser, 2013, Barrett and Stone, 2015] or plan recognition [Kaminka et al., 2002, Amir and Gal, 2013] to infer other agents' plans or parts of the environment model (e.g., transition and reward functions), these approaches still assume a known planning domain (i.e., known state space and actions in MDP frameworks, or known plan library in plan recognition approaches). In many distributed human teamwork settings, such plan models are rarely explicitly specified. For example, the complex health care teams we studied might agree on high-level treatment goals but never fully specify a long-term plan. The approach we present for ISLET problems does not rely on the complete plan knowledge assumption.

Chapter 4

A Personalized Change Awareness Mechanism for Collaborative Writing

Change awareness mechanisms [Tam and Greenberg, 2006, Dourish and Bellotti, 1992] are a prominent approach for supporting coordination in distributed teams. They assist team members in tracking each others' activities. By enhancing team members' knowledge of others' activities, they provide context for evaluating their own actions and ensuring they align with the team's activity as a whole [Dourish and Bellotti, 1992]. Change awareness mechanisms include change tracking (e.g., Microsoft Word's "track changes" feature, Google Docs' "suggesting" feature), Diff tools (e.g., comparing two versions of code on GitHub) and commenting features.

While some change awareness mechanisms allow team members to filter the information presented to them (e.g., to omit styling changes in a document, to show changes made by a particular collaborator), they do not reason about the *relevance* of such information to a particular team member. Team members therefore need to identify the information they should review and to notify their collaborators about changes they have made if they expect those changes to affect collaborators' activities. Consequently, team members often face one

or both of the following coordination challenges: (1) high coordination overhead as a result of information overload when too much information is shared, or (2) coordination failure due to lack of important information when too little information is shared or when relevant information cannot be found [Lowry et al., 2004, Steves et al., 2001, Pinelle and Gutwin, 2006, Amir et al., 2015].

In this chapter, we introduce the notion of *personalized change awareness* mechanisms. Such mechanisms reason about the structure of the collaborative activity to identify and share with team members only a subset of the change information that is deemed most relevant for their own activities. Personalized change awareness mechanisms have the potential to improve coordination in distributed loosely-coupled teamwork through simultaneously (1) lowering coordination overhead by reducing the total amount of change information each team member needs to review, and (2) improving the chances of coordination success by ensuring that each team member receives the information most relevant to her task.

We describe an implementation of a personalized change awareness mechanism for supporting collaborative writing that uses the MIP-DOI algorithm. When an author returns to a shared document, the mechanism highlights a subset of edits made by collaborators that are deemed most relevant to her, thus limiting the amount of information she needs to review prior to making her own writing contributions. We also describe extensions to MIP-DOI that make use of additional domain-specific knowledge in the context of writing (e.g., document structure).

Lastly, the chapter describes an evaluation of this personalized change awareness mechanism, demonstrating its benefits to team members. In the experiment, we compared personalized change awareness to two baselines: a change awareness mechanism that showed *all* of the changes, and a change awareness mechanism that showed the same number of changes as the personalized mechanism, but selected the changes to share at *random*. Our results show that both mechanisms that restricted the number of changes shown, *Personal-*

ized and *Random*, led to reduced participants' perceived workload and higher productivity compared to the condition in which all changes were shared, demonstrating the benefits of personalized change awareness with respect to reducing coordination overhead. They also show that personalized changes shared with participants were rated as more helpful than randomly chosen changes, and that the quality of the team's documents was higher with the personalized mechanism compared to the random mechanism, demonstrating the importance of sharing relevant changes.

4.1 A Personalized Change Awareness Mechanism for Collaborative Writing

The approach for personalizing change awareness in the context of collaborative writing comprises three steps, described in the consequent subsections: (1) mapping paragraphs across revisions and identifying changes to the document; (2) updating a MIP-Net representing the group activity, and (3) computing MIP-DOI to choose a subset of changes to share with a particular co-author. Steps (1) and (2) are executed at the end of each editing session. Step (3) is done when a particular author begins a new editing session.

For simplicity, we describe the approach assuming asynchronous editing, such that each revision is written by a single author, during a single editing session. Our approach, however, can also handle synchronous editing.

4.1.1 Mapping Paragraphs and Identifying Changes

The first step in our approach is to identify the paragraphs that changed when an author completes an editing session, and the extent of changes. This task is non-trivial because paragraphs can be moved, deleted and added over time. We use the method proposed

by Gehrmann et al. [2015] which computes pairwise similarities between paragraphs in the previous and current revisions to identify paragraphs that were moved.

Once paragraphs are matched, editing actions are identified by comparing the old and new versions of the same paragraph. If a paragraph has changed, the extent of change is computed using cosine similarity, and an *edit* action is added to the session's action list. New paragraphs are considered to have a change extent of 1 and an *add* action is included for each added paragraph. Similarly, each deleted paragraph results in a *delete* action in the session, and the change extent for these paragraphs is set to 1. The sequence of actions taken in the session and the identity of the author who took these actions (referred to as the *session's author*) are given as input to the MIP-Update procedure described next.

4.1.2 Updating a MIP-Net of the Collaborative Activity

The MIP update procedure (described in more detail in Section 3.2) iterates over the list of actions in the session, and updates the MIP-Net as follows. First, if the session's author is not yet represented in the MIP-Net, a new partner node is created. Similarly, a new node is added for each of the new paragraphs that were introduced in the session. Second, for each paragraph edited in the session, the weight on the edge connecting the node representing the session's author and the nodes representing the edited paragraphs are incremented. Third, the weight on edges connecting each pair of paragraphs that were edited in the session is incremented.

4.1.3 Using MIP-DOI to Reason About Information Sharing

At the beginning of a new editing session, the current MIP-Net is used to determine the changes to share with the author who began the new editing session, whom we refer to as the *current author*. MIP-DOI is applied to the full set of changes that were done since the

last editing session of the current author. This set of changes is extracted by combining all the editing actions from sessions that were done after the author's last editing session.

In the implementation of the personalized change awareness mechanism for collaborative writing, we extended the MIP-DOI algorithm to make use of additional information that is available in the writing domain in two ways. First, we consider the extent of the change made to the paragraph. This information is incorporated to distinguish between minor edits (e.g., typo fixes) and more substantial edits. We filter the set of changes considered for sharing to include only those changes for which the change exceeds a threshold of 0.05 (determined empirically using Wikipedia data).

Second, we consider the proximity of the author not only to the changed paragraphs, but also to the sections they appear in. By doing this, we utilize knowledge of the document structure to include more information about the authors' areas of responsibility and interests. While in many cases an author has not edited in the past the particular paragraph that has changed, knowing that the author has edited other paragraphs in the same section suggests possible interest in the content of the section. In sum, we consider all paragraphs that have been changed since the current author's last editing session and which exceed the change extent threshold. We rank them using the following DOI computation:

$$\text{DOI}(o \mid p) = \alpha \cdot API(n_o) + \beta \cdot D(n_o, n_p) + \gamma \cdot D(S_o, n_p)$$

where $D(S_o, n_p)$ is the proximity of the team member node to the nodes in the section that contains the modified paragraph, computed by averaging the proximity to each of the paragraphs in the section. Our implementation uses $\alpha = 0.1, \beta = 0.8, \gamma = 0.1$. We chose these values to give the highest weight to the proximity between the author and the paragraph, and based on empirical retrospective analysis of our own co-authored documents. We omitted the focus object component which was introduced in the DOI computation in Section 3.3 because

the personalized change awareness mechanism we implemented did not have information about the particular section authors intended to edit.

4.2 Experiment

To evaluate the personalized change awareness mechanism, we conducted an experiment in which participants collaboratively edited summaries of prominent news stories. We compared personalized change awareness to two baselines: one in which all changes were shown, and one in which only a subset of changes were shown (thus limiting the coordination overhead), but the changes were selected randomly. This design allowed us to separately investigate the impacts of the relevance and the quantity of the change information shared with each team member.

4.2.1 Participants

We recruited 18 undergraduate students to participate in the study (11 females, 7 males; aged 18 through 22). Participants received $40 for completing the study. Participants were divided into six teams of three.

4.2.2 Task

News summaries. Participants in the study collaboratively edited summaries of three news stories: Brexit, the coup attempt in Turkey, and the 2016 US presidential elections primaries. For each news story, participants were given an initial summary of the story (3000 words, 11 sections, 60 paragraphs), and were asked to expand the summary based on additional news items given to them in each session. We chose to start with an existing summary to make it non-trivial to identify all the edits made by co-authors from the start.

The initial news summaries were adapted from summary articles taken from major news sources.

In each of their editing sessions, participants were given a list of news items to add to the summary. A few examples of these news items are given in Table 4.1. We gave participants the news items so that they would focus their efforts on incorporating the information into an appropriate place and ensuring it aligned with other parts of the summary, rather than on composing new prose or synthesizing information. For the most part, this meant participants could simply paste each of the news item snippets into the summary (and update other places to avoid duplication of information or conflicts). In some cases participants changed some of the wording to make the text fit better within an existing paragraph. This choice led to a controlled setting in which the team performance did not depend on the writing abilities of team members but rather on their ability to maintain an up-to-date and coherent document.

The news items were taken from short news pieces published in Associated Press or Reuters, in chronological order, such that updates which appeared in the general media at the same time tended to appear together in our lists as well. Participants were instructed to ignore their knowledge of the news story and update the summary only based on the information provided to them.

This task setup mimicked the kind of work that is done in a real news desk, where one of the authors had worked. In the news desk, journalists update a shared document and need to avoid redundancies and conflicts under strict time constraints.

Editing teams. Participants worked in 3-person editing teams. A team size of 3 had the essential coordination complexity in terms of participants not being able to easily keep track of everybody's edits, but was small enough that the entire task could be completed in a reasonable time frame for an experiment (over 7–10 days).

Each team member was assigned an editing role. There were two types of editing roles, *specific* and *general*. Two members of each team were assigned *specific* editing roles. They

News story	Example news items
Brexit	UKIP leader Nigel Farage hailed it as the UK's "independence day", while Boris Johnson said the result would not mean "pulling up the drawbridge". (political editor)
	Britain's vote to leave the European Union has thrown financial markets into turmoil and means the U.S. Federal Reserve's ambitions for two rate rises this year have been placed on hold. (economics editor)
Turkey coup	Turkey's main opposition Republican People's Party (CHP) said the repose to a failed coup attempt must be conducted within the rule of law and that the plotters and those who helped them must be tried in the courts. (political editor)
	French President Francois Hollande said he expected there would be a period of repression in Turkey in the aftermath of a failed coup by some elements of the military. (foreign editor)
US Elections	More support for Clinton, as the race looks like it's over: Bernie Sanders endorses Democratic rival Hillary Clinton. 'She must become our next president', Sanders said in his statement. (democratic-party editor)
	U.S. Secret Service officials say security planning for the Republican national conventions took into account large-scale terrorism threats like the vehicle attack that occurred in France. (republican party editor)

Table 4.1: Examples of news items given to participants for each of the news stories.

were responsible for updating the summary with respect to a particular aspect of the news story. The *specific* roles were chosen based on the news story: political and economics editors for the Brexit story, political and foreign editors for the Turkey coup attempt story, and Republican-party and Democratic-party editors in the US elections story.

The *specific* editing roles meant that each of these editors' changes were, for the most part, focused on sections related to their primary area of responsibility. For example, the economics editor for Brexit made many of her edits in sections about implications of Brexit and membership in the European single market, while the political editor covered the process of electing a new prime minister. However, some news items were related to the areas covered by both editors (e.g., the election of a new prime minister affecting market performance), and therefore when one *specific* editor introduced such news items her edits often affected sections that were primarily edited by the other editor.

The third team member was assigned the *general* editor role, and was responsible for

adding news items that were not yet added by the *specific* editors, enhancing coherence and enforcing a 3000 word limit. The *general* editor thus had to keep track of everyone's edits and edited sections related to all aspects of the story.

4.2.3 Conditions

Our experiment included three conditions, differing in the change awareness mechanism used to share change information with team members. The three conditions were:

1. *All*: all of the changes made to the document since the participant's last edit were highlighted.
2. *Random*: a random subset of the changes was highlighted.
3. *Personalized*: a subset of the changes, considered most relevant to the participant using MIP-DOI was highlighted.

We included the *All* baseline to reflect current systems which show all the changes made to a document. This condition gives participants the possibility of reviewing all of the changes, but imposes more coordination overhead than mechanisms that limit the amount of shared information. Participants in the *Random* and *Personalized* were presented with a similar (small) number of changes to review. We included the *Random* condition as a second baseline to enable us to evaluate the impact of the personalized change selection independent of the number of changes that had to be reviewed.

With both the *Random* and *Personalized* conditions, up to 5 changed paragraphs were highlighted, with the limit that at most two-thirds of the changed paragraphs should be highlighted. We did not make the number of highlighted changes entirely conditioned on the total number of changes to avoid creating substantial differences in the number of highlighted changes between these two conditions. We chose to restrict the number of highlighted changes

to 5 as pilot studies determined the typical number of changes was 10-15, and we wanted to enforce substantial filtering.

In all conditions, the changes chosen to share were highlighted using the visualization style of Google Docs revision history view, which highlights modified text, marking deleted text with strikethrough. We used different highlighting colors for each of the editors.

Each team edited each of the three stories throughout the experiment, with each story being assigned with a different change awareness mechanism, thus enabling a within-subject comparison of the change awareness approaches. Within each group, the assignment of a condition to a story persisted throughout (e.g., in group 1 *Personalized* was used for Brexit, *Random* for Turkey and *All* for elections; in group 2 *All* was used for Brexit, *Personalized* for Turkey and *Random* for elections). This enabled us to evaluate the final summary that was produced when a particular change awareness mechanism was used. The ordering of the conditions and the assignment of conditions to stories across groups was counterbalanced (we had 6 groups, covering the 6 possible combinations of condition-story assignments).

4.2.4 Procedure

We designed the procedure to reflect extended-duration collaborative writing and to control the amount of information and number of changes that were introduced in each revision. To this end, we had four rounds of editing, with each participant completing one editing session in each round. To maintain a similar number of changes between each of the editors' sessions, the sessions were done in turns, with each editor returning to the document after the two other editors completed their sessions. The *general* editor was the last to edit in each round. Editing sessions were done remotely and were typically conducted one day apart from each other for each of the editors. The overall duration of the study for each team was 7–10 days during which all editors completed the four rounds of editing sessions, where in each of the rounds they edited all three stories. Due to scheduling constraints, one of the

Activity	Time
Review news story 1	Round 1: 10 minutes; rounds 2-4: 30 seconds
Edit news story 1	12 minutes
Questionnaire news story 1	2 minutes
Review news story 2	Round 1: 10 minutes; rounds 2-4: 30 seconds
Edit news story 2	12 minutes
Questionnaire news story 2	2 minutes
Review news story 3	Round 1: 10 minutes; rounds 2-4: 30 seconds
Edit news story 3	12 minutes
Questionnaire news story 3	2 minutes
Final questionnaire	Only at last round, 10 minutes

Table 4.2: The activities in each of the editing sessions.

groups completed only three editing rounds.

The duration of each editing session was limited to 12 minutes. During each session, the *specific* editors were given 10 items to add to the story. The length of the editing sessions and the number of news items were determined based on pilot studies, with the goal of creating a task that would be hard to complete in the given time frame, while maintaining a reasonable session length (up to an hour for completing the task for all three summaries).

Participants were not allowed to communicate with each other or to use the commenting feature on the document. While co-authors in real-world settings have many ways to communicate, we did not include any in the study because we needed a controlled environment in which change awareness of team members did not depend on the particular communication strategies of team members and was manipulated only by our experimental conditions.

Each editing session consisted of a cycle of three stages for each of the news stories, summarized in Table 4.2: (1) reviewing the current summary, (2) editing the summary, and (3) answering a questionnaire about their subjective experience.

In the **reviewing stage**, participants were first given time to read the current summary without ability to edit. In the first session, participants were given 10 minutes to read the current summary. In rounds 2–4, they were given only 30 seconds to review changes made

to the document by their co-authors. The purpose of the short reviewing period was to draw participants' attention to the fact that their co-authors' had made changes. The time provided was intentionally short so that participants would not be able to review all changes before beginning their own editing session. Participants were still able to review changes later while editing the document.

In the **editing stage**, *specific* editors were given 12 minutes to incorporate 10 news items assigned to them based on their roles. They were instructed to add as many of the items as they could, while maintaining coherence (e.g., adding the information in an appropriate place) and avoiding conflicts and redundancies. For instance, in Brexit, when adding information about Boris Johnson deciding not to run for prime minister, the editor would also need to remove information suggesting that Johnson is the leading candidate in the race. During the editing stage, participants could still review the changes made by other team members. For the first three groups who participated in the study, the highlighted changes were shown on a separate, non-editable document, and participants could switch between editing and reviewing by clicking a button. Because many participants in these first groups commented that they would have preferred to have the changes highlighted on the document they were editing, for the remaining 3 groups the highlighted changes appeared on the document they were editing; they were instructed that they could remove the highlighting if they wished[1].

General editors were given the news item lists of both *specific* editors, and were asked to add any missing information and make any other required edits to make the summary coherent. They were also asked to maintain a 3000 word limit, which meant they had to edit even if the *specific* editors managed to add all the news items during their sessions.

The editing stage was followed by a **questionnaire** about their experience which is described below. The process of reviewing, editing and answering the questionnaire was

[1]We did not find differences between the earlier and later groups and therefore analyzed them together despite the difference in the way changes were presented.

repeated three times (once for each news story).

After completing their last session, participants were asked to answer a final questionnaire comparing the different change awareness mechanisms (referred to as "highlighting methods" when communicating with participants). Participants were provided access to the summaries with the highlighted changes and were asked (for each summary separately) to identify up to two highlighted changes that were relevant and helpful for them, and up to two highlighted changes that were irrelevant and unhelpful. They were asked to comment on the ways in which the relevant changes helped them. Participants were also asked to rank the highlighting methods based on their preferences. Because participants did not know which method was used for each news story, they were provided the associated story for each method (e.g. "method 1 (Brexit)").

4.2.5 Design and Analysis

We used a within-subject design: each team edited one summary in each of the conditions. Each participant held a similar role (*specific* or *general* editor) in each of the three summary tasks.

Our dependent measures consisted of both objective and subjective measures related to participants' performance and experience. Participants provided subjective responses about the following aspects of the task at the end of each editing session for each summary in each round. All questions are shown in Table 4.3:

- Teamwork (e.g., "I was able to build on my co-authors' work").
- Helpfulness of the shared changes (e.g., "seeing the highlighted changes helped me ensure my edits align with others' edits").
- Workload assessment using the NASA TLX scale [Hart, 2006]. (For example, "How mentally demanding was the task"; we omitted the physical demand question as it was

irrelevant for our task).

For teamwork and helpfulness of changes statements, participants rated their level of agreement on a 7-point Likert scale (1 = "strongly disagree", 7 = "strongly agree"). Workload questions (NASA TLX) were rated on a 7-point Likert scale (1 = "very low", 7 = "very high").

In the analysis of perceived workload, we considered the sum of all 5 items in the NASA TLX scale, which is common in the analysis of TLX measures [Hart, 2006]. To create a consistent scale, we take the complementing value of the "success" measure, which is the only positive item in the scale.

Our dependent measures also included the following performance measures:

- Coverage: the number of news items that were added to the summary.
- Quality: the quality of the final document in terms of consistency and coherency.

We measured coverage at the session level (i.e., the specific contributions made by a single participant in a single session). Assessment of coverage was objective (was a particular factual item incorporated into the summary or not) and was done by the authors. We measure coverage only for *specific* editors because the number of news items incorporated by *general* editors largely depended on the number of changes introduced by their team members.

We assessed the quality metrics for the final summary, which reflects the final product of the teamwork. Assessing the quality was done by comparing the final document to the initial document, and rating each paragraph that was added along the following dimensions (each rated on a scale from 1=poor to 3=good): (1) whether the text is redundant with other parts of the summary or conflicts with other information in the summary; (2) whether the paragraph is coherent with the remainder of the section; (3) whether it appeared in an appropriate section based on the section titles (e.g., information about May being elected should appear in the section about the race for the new prime minister), and (4) whether it

appeared close to other related information. The reason for considering (3) and (4) separately was that in some cases an editor introduced text under an inappropriate section title, and later another editor placed related news updates nearby. If so, we wanted to penalize the new edit in terms of section location, yet acknowledge that it was rightfully located next to relevant information.

The quality assessment was done by two judges, with each summary being evaluated by a single judge, blind to condition. Both judges used the same assessment rubric, and they calibrated their ratings by evaluating one of the summaries independently and then comparing their evaluations. Their initial agreement was good (Krippendorff's alpha of 0.7) and they resolved any disagreements to be more aligned moving forward. One of the judges was an author who was not involved in running the study and who had worked in the past as a news editor. The other judge was an undergraduate student in the humanities. Because we were interested in comparing the performance within each of the groups, we were more concerned with having consistent rating for the three summaries that the same group edited. Therefore, each of the judges evaluated all 3 summaries (Brexit, Turkey coup, US elections) of 3 of the groups.

Analyses. Our analysis of subjective measures and of the objective coverage measure is done on the data from the last round of the study. We focused on the last round as we expected the personalized mechanism to take time to learn about participants' areas of responsibility. The quality assessment was done on the final summary as a whole, and therefore reflects the teamwork throughout all rounds.

We analyzed the Likert-scale items and the number of items added (between 1–10) using ordinal logistic regression. The main effects we considered were the condition and the editor type, and we controlled for the topic of the news summary. We also included the participant id (unique identifier of each participant) as a covariate as our study was a within-subject design. Each participant id appeared three times in each analysis, once for each of the

conditions. Thus, our model included the condition {All, Personalized, Random}, editor type {Specific, General}, topic {Brexit, Turkey coup, US elections}, participant id {1,...,18}.

Ordinal logistic regression estimates the likelihood of obtaining a *higher* value of a dependent measure given a change in the values of the independent variables. The fitted model can be meaningfully interpreted by considering the *odds ratio* (OR) values of the regression coefficients of the independent variables[2], which we report in the Results section. For example, when estimating the effect of the study condition on participants' workload, the odds ratio value of *All* vs. *Personalized* was 8.8. This odds ratio value is interpreted as follows: all else being equal (i.e., same topic, participant and editor type), the odds of reporting a workload value that is *higher* than k (for any possible value of k) in the *All* condition are 8.8 times as large as the odds of reporting a value higher than k in the *Personalized* condition. That is, participants are much more likely to experience higher levels of workload when seeing all changes than with personalized change awareness.

To make interpretation easier, we always report odds ratio values greater than one. To do this, we take the multiplicative inverse value of odds ratios lower than 1, and interpret them as the odds ratio for obtaining a *lower or equal* value of the dependent measure (rather than a higher value). For example, if we get an odds ratio of 0.5 for the likelihood of obtaining a higher value, we convert it to an odds ratio of 5 for obtaining a lower or equal value. Odds ratios can be interpreted as effect size, similar to Cohen's d. Values between 1.5 and 3 are interpreted as a small effect, between 3 and 5 as medium, and above 5 as large [Borenstein et al., 2009, Chen et al., 2010].

When we observed a significant main effect, we re-ran the model to compute the significance and effect sizes (odds ratios) of pairwise differences.

To account for multiple hypotheses testing, we adjusted p-values with the Holm's sequen-

[2]Odds ratio values are computed by exponentiating the regression coefficients, which estimate log odds ratios.

tially rejective Bonferroni procedure, which introduces fewer Type II errors than the simple Bonferroni correction [Holm, 1979, Shaffer, 1995]. We did this separately for subjective and objective measures. In post-hoc analyses we adjusted the p-values of pairwise comparisons with the simple Bonferroni correction.

Some Likert-scale items were intended to measure the same construct. We combined participants' responses to those items if they had an acceptable Cronbach's alpha ($\alpha > 0.7$). This resulted in combining the responses to the items "seeing the highlighted changes helped me decide where to make my edits" and "seeing the highlighted changes helped me ensure my edits align with others' edits", and combining the responses to the items "I was overwhelmed by the changes to the document's content" and "Keeping track of my co-authors' edits was difficult."

The quality measures were also analyzed using ordinal logistic regression (scales of 1–3), and included as effects the condition, topic of the summary and the group that produced the summary[3]. We combined the four quality ratings (Cronbach's $\alpha = 0.83$) into a single quality measure.

Participants' final ranking of the change awareness mechanisms was analyzed using the nonparametric Friedman test.

4.3 Results

Table 4.3 summarizes the results of the study. While ordinal logistic regression does not directly compare means, we show the mean values to illustrate the results. When a statistically significant difference was found between two conditions, we report the *odds ratio* values that quantify the difference between conditions.

[3]We did not find an effect of the judge on the ratings, so did not include it in our model.

	Measure	Personalized (Mean)	Random (Mean)	All (Mean)	Adjusted p-values	Significant pairwise differences
Workload	TLX (5–35, 5 is best)	18.2	**17.8**	21.6	**0.001**	P-A (OR: 8.88); R-A (OR: 9.04)
Teamwork measures	I was able to build on my co-authors' work with my edits. (1 – 7, 7 is best)	3.76	3.65	**4.12**	0.46	-
	I was overwhelmed by the changes to the document's content; Keeping track of my co-authors' edits was difficult. (2–14, 2 is best)	7.29	**6.82**	7.94	**0.0474**	R-A (OR: 7.4)
Helpfulness of changes	Seeing the changes made by others helped me decide where to make my edits ; Seeing the changes made by others helped me ensure my edits align with others' edits. (2–14, 14 is best)	**7.59**	7	7.47	0.49	-
	Seeing the changes made by others did not help me make my own edits . (1–7, 1 is best)	**3.12**	4.29	4.06	**0.024**	P-A (OR: 6.5); P-R (OR: 6.6)
Preferences	Final preference ranking (1–3, 1 is best)	**1.83**	2.06	2.06	0.77	-
Performance measures	Coverage (1–10, 10 is best)	**6.92**	**6.92**	5.67	**0.04**	P-A (OR: 4.8); R-A (OR: 6.5)
	Quality (1–3, 3 is best)	**2.8**	2.75	2.78	**0.04**	P-R (OR: 1.69)

Table 4.3: The means for each of the measures, and the Holm-Bonferroni adjusted p-value of the test for the significance of the condition effect. In cases where the differences were significant, the last column shows which pairwise comparisons of conditions were statistically significant (P-A indicates a significant differences between *Personalized* and *All*, R-A indicates a significant differences between *Random* and *All* and P-R indicates a significant differences between *Personalized* and *Random*), and the odds ratio (OR) of the difference.

4.3.1 Workload

We observed a significant main effect of condition on subjective workload (${\chi^2}_{(2,N=54)} = 14.8, p = 0.001$). Participants experienced significantly higher workload when seeing changes in the *All* condition compared to either *MIP* ($OR = 8.88, \chi^2_{(1,N=54)} = 10.3, p = 0.006$) or *Random* ($OR = 9.04, \chi^2_{(1,N=54)} = 10.6, p = 0.003$) conditions. The effects sizes for both pairwise comparisons were large ($OR > 5$).

4.3.2 Teamwork-Related Items

We observed a significant main effect of condition on participants' difficulty of keeping rack with others' work and their feeling of being overwhelmed by changes (${\chi^2}_{(2,N=54)} = 8.3, p < 0.05$). Participants were more likely to report greater difficulty of keeping track of others' changes in the *All* condition compared to *Random* and *Personalized*. The only statistically significant pairwise comparison was between *All* and *Random* ($OR = 7.4, \chi^2_{(1,N=54)} = 7.5, p = 0.02$).

We did not find significant differences in participants' reported ability to build on their co-authors work.

4.3.3 Helpfulness of the Shared Changes

With respect to the helpfulness of changes, we found a significant main effect of condition (${\chi^2}_{(2,N=54)} = 10.2, p = 0.024$). The shared changes were found more helpful in the *Personalized* condition compared to both the *All* ($OR = 6.5, \chi^2_{(1,N=54)} = 7.2, p = 0.02$) and *Random* ($OR = 6.6, \chi^2_{(1,N=54)} = 7.3, p = 0.02$) conditions.

To better understand the ways in which seeing changes helped participants accomplish their editing tasks, we examined participants' open-ended responses. At the end of the final editing session, participants were asked to copy up to 2 highlighted changes that were relevant

and to explain how each of the relevant highlighted changes helped them make their edits. A common response among *specific* editors was that seeing others' edits helped them navigate the document, and that sometimes the highlighted edits addressed topics similar to some of their news items, and thus seeing where they were located helped them decide where to add new information. For example, one participants wrote "It helped me know where to place a couple of my items and allowed me to check that certain info was updated", while another participant commented "I saw them [the changes] as a context for the news items I received later."

General editors commented both about the helpfulness in terms of deciding where to place items and knowing which items were already added by their co-authors, e.g. "The first edit was helpful because it informed me where in the document a lot of the political edits were made (especially concerning Erdogan's 'crackdown') and seeing it also made me quickly aware that I did not have to add that piece of information, which was included in one of the lists."

4.3.4 Preference Rankings

The *Personalized* mechanism was ranked best ($M = 1.83$), followed by *All* and *Random* (both with $M = 2.06$), but this result was not statistically significant ($\chi^2_{(2,N=18))} = 0.5352, p = 0.765$). While about the same number of participants rated *All* and *Personalized* as most preferred (8 and 7, respectively), the ranking of *All* was more bimodal: 9 of the 10 participants who did not choose *All* as most preferred, ranked it as *least* preferred. In contrast, 7 of the 11 participants who did not rank *Personalized* first, ranked it second. Random was ranked first only by 3 participants, but was ranked second more often than *All*.

Interestingly, we did not find different patterns in the ranking by *general* and *specific* editors. We expected that *general* editors may prefer to see more changes because their role required more awareness of the other editors' work, but observed that half of them ranked

All as most preferred while the other half ranked it as least preferred.

Participants were asked to explain their rankings. One participant who ranked *Personalized* as most preferred commented, "since the edits were very concentrated for the Turkey summary [*Personalized* condition], I ranked it as my most preferred method." Another participant who preferred *Personalized* commented about the summary which was associated with the *All* condition that "When the edits were spread further across the summary, it was harder to keep track of them all". Some participants who preferred the *All* condition mentioned that they liked seeing more changes, e.g., "I find seeing all of the co-editors' changes helpful in determining what to look for when correcting/adding information."

4.3.5 Participants' Performance

We observed a statistically significant effect of condition on the number of news items added by participants (${\chi^2}_{(2,N=36)} = 6.3, p = 0.04$). Participants added fewer items in the *All* condition compared to both *Personalized* ($OR = 4.8, \chi^2_{(1,N=36)} = 3.96, p = 0.04$) or *Random* ($OR = 6.5, \chi^2_{(1,N=36)} = 4.86, p = 0.03$).

We also found a statistically significant effect of condition on the quality of the final summaries (${\chi^2}_{(2,N=719)} = 7.72, p = 0.04$). Summaries produced in the *Personalized* condition were significantly more likely to have higher quality rating than summaries produced in the *Random* condition ($OR = 1.69, {\chi^2}_{(1,N=719)} = 7.56, p = 0.006$). Other pairwise differences between conditions were not statistically significant.

4.3.6 Personalized Sharing of Change Information

To better understand the way in which personalization was reflected by the changes shared in the *Personalized* condition, we examined the MIP-Nets modeling the teamwork in the different groups. We illustrate the personalization of change sharing with examples

of changes that were deemed relevant and irrelevant to different editors as this summary evolved.

Figure 4.1 shows the MIP-Net at the end of all rounds of editing of the Brexit summary produced by group 6. Author nodes are shown in red, with 'g', 'p' and 'e' representing the general, political and economics editors correspondingly. To make the network more compact, we collapse all nodes in a section to a single "section node" (blue nodes). The weights on edges connecting author and section nodes were computed by averaging the proximity of the author to each of the paragraphs in the section. The weights on edges connecting two sections were computed by averaging the proximity of all pairs of paragraphs in both sections. The intensity of edges connecting pairs of nodes reflects the weight of the edge (darker edges reflect higher weights).

The loose-coupling of the task can be seen in the MIP-Net, in that each editor node is most strongly connected to nodes representing different sections. The political editor is most strongly connected with nodes representing sections 1 ("Supporters and opposers of leaving the EU"), 4 ("Is the kingdom still united?") and 5 ("A new conservative Prime Minister to be announced"), which focus on political aspects of Brexit. In contrast, the economics editor is most strongly connected with the nodes representing sections 6 ("Britain's economy awaits Brexit's aftermath") and 8 ("British taxpayer's money to be directed to NHS?") which focus on economical aspects of Brexit, and section 10 ("Other implications of Brexit"), which evolved to include economics related information during this group's edits (e.g., it included a news item about Obama warning against financial hysteria following the vote). While the focus of the *specific* editors was on different sections, note that the two *specific* editors also overlapped in editing other sections, hence their activities were not entirely decoupled. The proximity of the *general* editor to the various sections was more evenly distributed, which is expected as that editor was responsible for the overall summary.

By capturing each editor's unique focus, the MIP-Net enables personalized assessment of

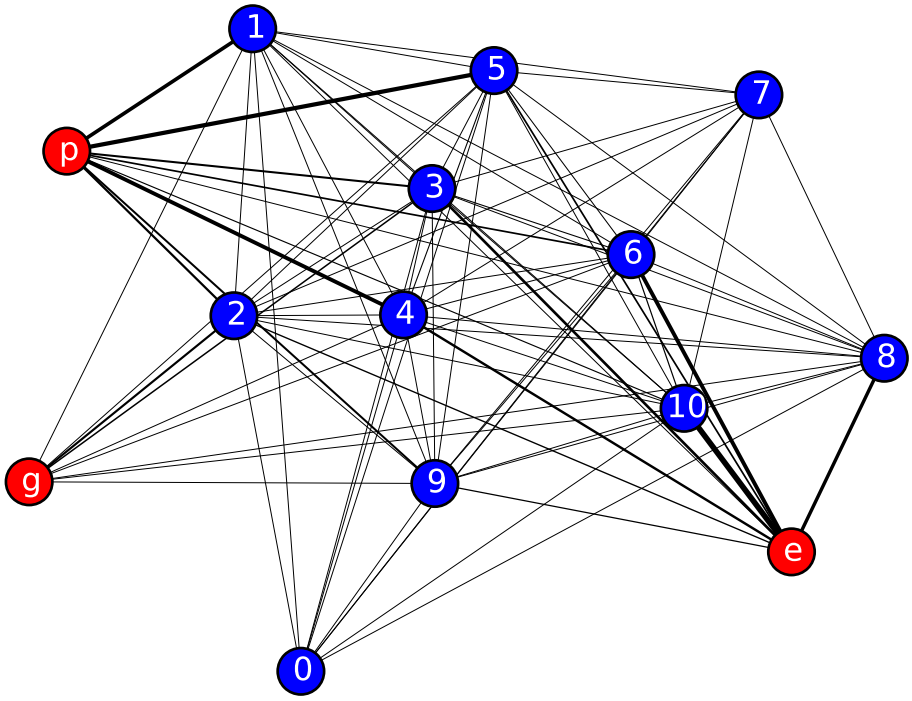

Figure 4.1: The final MIP-Net for the Brexit summary written by group 6, shown at the section granularity level. Author nodes are shown in red ('e': economics editor, 'p': political editor, 'g': general editor), section nodes are shown in blue. The intensity of edges corresponds to their weights (darker edges have higher weight).

relevance of edits to them. To illustrate, we consider the relevance assessments for two of the news items which were added to the summary by the *general* editor during her second round of editing:

1. *"Morgan Stanley sources said that it had started the process of moving about 2,000 staff based in London to either Dublin or Frankfurt. Ahead of the vote, the president of the investment bank, Colm Kelleher, told Bloomberg that Brexit would be 'the most consequential thing that we've ever seen since the war"'*
2. *"Germany's foreign minister says he hopes new British Foreign Secretary will want to maintain a 'sensible relationship' with the European Union."*

The first item was considered relevant for the *economics* editor. It had a high degree of interest because of its proximity to edits made by the economics editor; it appeared in the section "Britain's economy awaits Brexit's aftermath" (blue node #6 in Figure 4.1) which was substantially edited by the economics editor in previous editing sessions. This paragraph was rated much lower in terms of relevance to the political editor, as its proximity to that editor's edits was much lower. In contrast, the second item, which appeared in a section that discussed more of the political aspects of Brexit was considered relevant to the *political* editor, but not to the economics editor.

These two news items exemplify cases in which edits were fairly clearly related to either the economic or political aspects of the Brexit summary. However, some aspects of the story required more tight coupling of edits. For example, the section "Britain to stay in the single market?" (blue node #3) touched both on economy and politics related implications of Brexit, and therefore both editors had relatively high proximity to this section. Similarly, although the section "Britain's economy awaits Brexit's aftermath" (blue node #6) was primarily edited by the economics editor, the political editor occasionally edited it too, for instance when adding information about the G-20 summit which was already discussed in

that section. In such cases, a news item added by one of the specific editors would have a relatively high degree of interest to the other specific editor, despite the differences in their areas of responsibility. We note that such dependencies take longer for the algorithm to learn about.

4.4 Related Work

Change awareness mechanisms draw team members' attention to changes others have made to a shared artifact by marking those changes in some way. Tam & Greenberg [Tam and Greenberg, 2006] provide a framework that describes the types of questions that can be answered by change awareness mechanisms, such as where changes have been made, by whom and when. The framework also describes information elements based on which a system could answer these questions, including edit history, location history and authorship history.

Many current commercial tools provide CA mechanisms: Word's track changes, Google Docs suggest mode and revision histories, GitHub commits, Wikipedia revision history are examples. These tools, however, typically show users *all* of the changes that were made by their collaborators, which can lead to information overload.

Some prior CA approaches provide team members with change filtering options. For example, PastDraw provides filtering options such as choosing the types of changes to show (e.g., deletions, additions) and for which types of objects to display changes [Tam and Greenberg, 2006]. Prinz et al. proposed "anticipative" CA mechanisms [Prinz et al., 2010], which allow team members to specify ahead of time the changes they would like to be notified about (for example, letting a user know when a document was opened by someone). In the context of collaborative writing, flexible Diff-ing [Neuwirth et al., 1992] provides users with ability to filter changes based on features such as the granularity of the edits.

While these approaches can help reduce information overload, they require manual input

from the users in selecting the changes to show. In contrast, the intelligent personalized change awareness mechanism we developed reasons about the relevance of changes to users and does not require that users manually specify the types of changes they are interested in. We use the same types of information elements described by Tam & Greenberg, in particular the *edit history*, but in a different way: rather than using the edit history only to extract change information, it is used by the system to learn about the interests of team members, the task allocation among collaborators and the dependencies between the tasks.

Most closely related to our approach are methods for filtering notifications in software development settings. The NeedFeed system [Padhye et al., 2014], for example, models code relevance for developers by analyzing "touch histories" (which classes a developer has modified) and using more complex history-based classifiers. Similarly, Holmes and Walker [2010] proposed a recommendation approach to filter notifications about change events based on deployment dependencies. Their approach uses various code features and code ownership analysis. Omoronyia et al. [2009] developed a system that aims to enhance collaboration awareness by showing tasks, developers and artifacts that are considered relevant to a user given their current context.

These software development approaches are similar in spirit to our approach for personalized change awareness, but they rely heavily on the explicit structure that is available in software programs (e.g., class dependencies, code documentation, method call graphs, use cases). In many domains, however, there is much less explicit structural information. For instance, in writing there may be a hierarchical section structure, but the dependencies between sections are much less apparent than in code. In addition, the evaluations of these systems assessed the relevance of the identified changes based on code revision histories, but did not directly evaluate the benefits that these approaches provide to the teamwork.

Other relevant prior work includes visualization tools that support distributed software development by creating visual representations of software artifacts and development activi-

ties [Froehlich and Dourish, 2004, de Souza et al., 2007, Jakobsen et al., 2009, De Souza et al., 2005, Gutwin et al., 2004]. In contrast to the personalized CA approach, these visualizations help users be aware of other artifacts and people who might be affected by their work rather than trying to draw their attention to relevant changes. Importantly, they do not create *personalized* views. Similarly to the systems described above, they also rely on the available structural information in software development.

Chapter 5

GoalKeeper: Supporting Care Plan Management

Integrated team-based care plans provide a shared context for complex care teams and improve care coordination. However, current healthcare systems do not support the use of such care plans. Electronic medical records support linear processes rather than dynamic coordination processes [O'Malley et al., 2010]. They do not enable team members to define partial plans that can be expanded and revised over time. Our study of care teams (described in Chapter 2) revealed that team-based care plans require continuous revisions as a result of changes in patients' conditions, and thus quickly become obsolete if they are not updated. In addition, collecting information about patients' progress toward goals in a consistent manner was also found to be challenging.

This chapter describes the design and implementation of GoalKeeper, a system for supporting the use of team-based care plans for complex patients. GoalKeeper enables families and their care providers to define care goals for the child, specify actions that need to be taken to make progress toward these goals. It further enables team members, in particular the patient's family, to track progress toward goals by recording status updates about the

child's condition. The chapter also presents an initial pilot study in which two families used GoalKeeper to create and track care plans. GoalKeeper was found helpful for organizing care activities and for making apparent when goals were accomplished or when progress was not achieved, prompting users to take actions accordingly. However, study participants found it challenging to continuously record information given their many other daily activities, and also reported negative emotional impact of observing lack of progress toward goals.

5.1 Designing a System to Support Care Plan Use

The design of GoalKeeper was informed by our study of care teams. In addition to the findings described in Section 2.2, we also elicited caregivers' thought about technology for supporting the use of care plans. At the end of the interviews we conducted with care team members, we asked interviewees to reflect on ways technology might support care teams in the use of care plans. At first, these conversations were entirely open-ended. Later, as patterns began to emerge, we would bring sketches and mockups of designs based on the insights from earlier interviews (e.g., the one in Figure 5.1). These initial designs were very rough, and the discussions focused on the content of the system rather than interface design. The designs included a list of goals, each with a description, responsible caregivers, status updates showing progress towards the goal, and a list of pending actions. In the later conversations, we started by asking interviewees what features they would like to see in such a system before revealing the mockups, then showed them the mockups and discussed what they found useful in them and what was missing. In many cases, showing the mockups spurred additional reflections on problems in complex care, concerns about introducing such a system and expected benefits of the system.

Before discussing a care plan support system with parents, we asked parents about the information they currently track and the tools they use to do so. Parents reported keeping

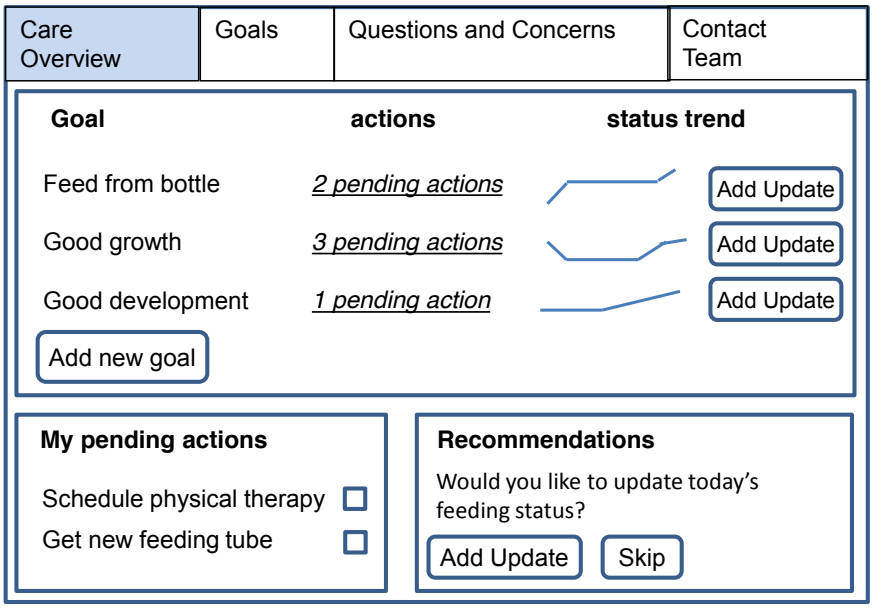

Figure 5.1: An example screen mockup of GoalKeeper

track of a variety of information, such as food pump settings, medication doses, ventilation settings, results of blood tests and contact details of care providers. They use a variety of tools to track this information, including online tools (e.g., google calendar, notes applications) and paper lists (e.g., lists of medications or contact details of providers). One parent reported using online tools to store EMR notes, maintain a provider contact list and the appointment schedule, taking notes and tracking height and weight. Another parent reported maintaining a handwritten medication lists, and mobile apps for taking notes and tracking height and weight. Another parent reported saving the paper summaries she receives from the neurologists and test results documents. Many of the providers we interviewed said that parents often come with big binders where they store their documents. Overall, parents did not have a unified place where they could collect and track all of the information, highlighting the potential for improved technology support.

With respect to the mockups presented, parents found the goal-oriented approach of the design appealing. One said he particularly liked the idea of being able to "tag" a goal to multiple providers and then sort goals based on tag to prepare for an appointment. Another

parent liked being able to track the status of her child: "this would have been really helpful when my daughter was in the hospital, especially being able to keep track of things." Parents added that they would like to have a place to write a short medical summary of their child's condition, which they would update every couple of months. They also wanted to have a place to document questions and issues they need to remember to bring up in future visits.

Parents differed in how they envisioned using the system. One parent said she would want her family to specify goals and have providers define actions required to achieve them. She wanted providers to be accountable for their actions. Another parent wanted the system to remind her of actions she is required to do rather than providers' actions.

Given their experiences with tracking information, parents were concerned about the work involved in providing status updates reporting the child's status with respect to a goal (e.g., number of feeds each day) and whether the effort would pay off. They were concerned that providers would spend too much time during the visit looking at the computer system rather than focusing on their child. They were also concerned that providers would not look at the information between visits, especially the specialists. They emphasized the need to make the system customizable and not overwhelming by having an option to hide more detailed information.

As did parents, some providers were concerned about the burden of using another system: "If there are three different systems like this and we're seeing patients with each different system and every time I go in, there's a new system and I'm also having to document things from the chart, it's a no go. So I think the simplest format of this is actually providing a patient facing system, because the doctors have their notes and so to a certain degree, they're going to have their system."

Despite their concerns about the burden of using an additional system, providers saw possible significant benefits of a care plan support system. One specialist thought that having a system owned by the parents where they can keep track of their goals and what providers

said to them would provide feedback for the provider to see what they have heard at the encounter and create an opportunity to interact. She also thought that if family-oriented goals were set by the family it would empower them. Another specialist said, "especially because so many parents want a way to sort of synthesize that information and I think they would really benefit from that [...] And if you have the parents synthesize it and putting it in a language that they presumably would understand and then you sort of see it translated in real time, I think that's valuable." That specialist suggested that taking a multidisciplinary view in such a system would be useful, as it will be more similar to how parents view the information, as opposed to the medical organ system based organization: "If you explain it for the family, neurology and pulmonary, it does provide a little simplicity in terms of they know what questions to ask you, but that may not always be doing them a service, because the way we group things is a little artificial to understand." He added that synthesizing information can help physicians see what other providers are doing and thinking.

Therapists thought tracking goals with a computer system would be more efficient than tracking currently is, "if you've met a goal you could write a new one right away. You wouldn't have to wait." One therapist was concerned that it could be a negative thing: "They're [the parents are] seeing all these curves going down". However, she said that noticing the decline could raise a "red-flag" in time and draw the attention of the care team to the problem.

5.2 The GoalKeeper System

Based on our study findings, we have designed a system, called GoalKeeper, to support the creation, monitoring and revision of care plans. GoalKeeper is designed to be "owned" by parents, rather than by care providers, for the following reasons: (1) The team is built around the family who care for the patient at all times. (2) Giving the family control of the care plan can empower them and help them interact with professional members of the care

team. (3) Providers are already overwhelmed by demands for them to input information into existing systems. (4) This choice preserves privacy by enabling parents to decide who can access GoalKeeper's information about their child.

In accordance with the Packard report recommendations for care plans and based on the SharedPlans specifications for collaborative activities, the main constructs in GoalKeeper are *goals*, *actions* to be taken towards accomplishing the goals, and *status updates* for monitoring progress with respect to goals. We note that these constructs are relatively simple compared with AI planning representations such as hierarchical task networks [Erol et al., 1994] which include more complex hierarchies. We chose to not use a deeper hierarchy of goals and actions based on discussions with providers, as the simpler representation matched the level of information that providers need to share and they were concerened that more complex representations will introduce unneccessary complexity and might make the system less usable.

GoalKeeper provides a platform for team members to create, share, monitor and revise the care plan. Team members can set care goals for the patient and specify actions that should be taken to achieve each a particular goal. They can specify the set of team members that is associated with each goal. The main view of GoalKeeper, shown in Figure 5.2 lists the currently active goals, taken from the care plan example provided in Section 2.1. For each goal, the following information is shown: a general description of the goal (e.g., "get Alex into day care in the next two months"), a graph showing progress toward the goal based on recent status updates, a list of the care team members associated with the goal and a list of pending actions associated with the goal. New goals can be added using the form shown on the right side of Figure 5.2.

Team members can view more details about a particular goal and add actions and status updates by going to that goal's page (by clicking on the title of the goal or the "view this goal" button). Figure 5.3 shows the page for the "take milk by mouth" goal. The top of the page

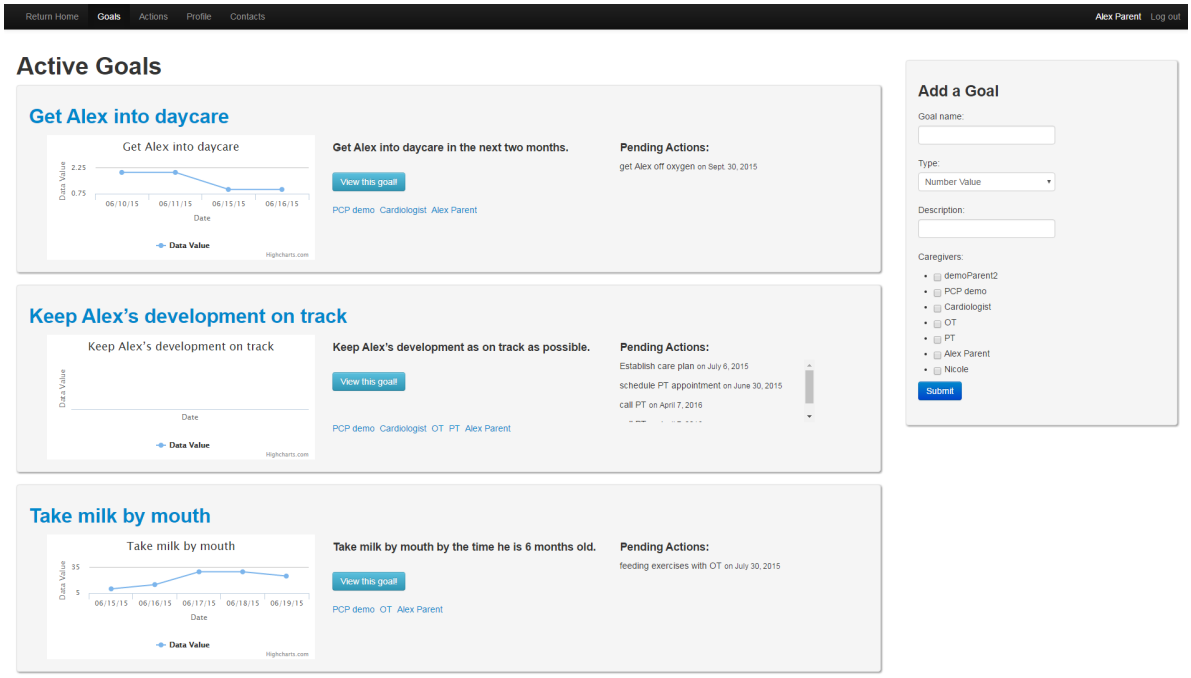

Figure 5.2: The "Goals" page. Displays the current goals for the patient.

shows a graph based on the status updates recorded for the goal, in this case, milk intake. Below the graph, the left side lists the recent status updates along with any notes added by the parents; the right side shows a list of the pending and recently completed actions associated with the goal. Team members can add status updates and actions using the forms at the bottom of the page.

The "Actions" page, shown in Figure 5.4, aggregates the actions from all the active goals. The goal to which each action is associated is listed along with the description of the goal. Users can mark actions as completed or remove them, as well add new actions using the form on the right side.

In addition to specifying and revising the main components of the care plan, i.e., the goals actions and status updates, GoalKeeper also includes a Profile page (shown in Figure 5.5) which provides a high-level summary of the patient's condition and a Contact page (shown in Figure 5.6) that lists the members of the care team.

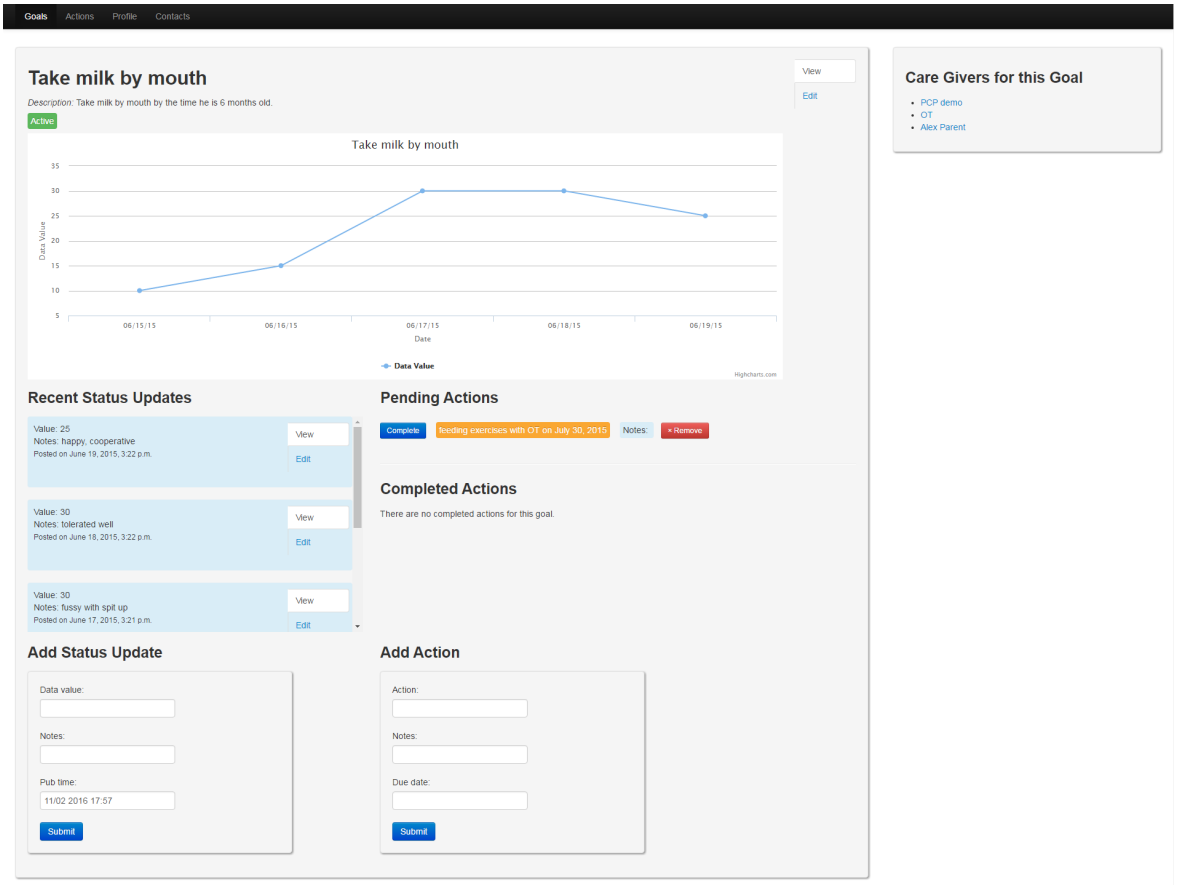

Figure 5.3: The individual goal page for the "take milk by mouth" goal. The page shows status updates and actions related to the goal, and a list of the caregivers involved in achieving the goal.

5.3 Pilot Study

We conducted a preliminary study of GoalKeeper to evaluate its use and to inform future designs. This initial study focused on families' use of GoalKeeper and did not directly involve other care providers.

5.3.1 Participants

We recruited two families through "Mass Family Voices", an organization that supports families of children with special healthcare needs. For one of the families, the user of the

Figure 5.4: The "Actions" page. Displays the pending and recently completed actions across all goals.

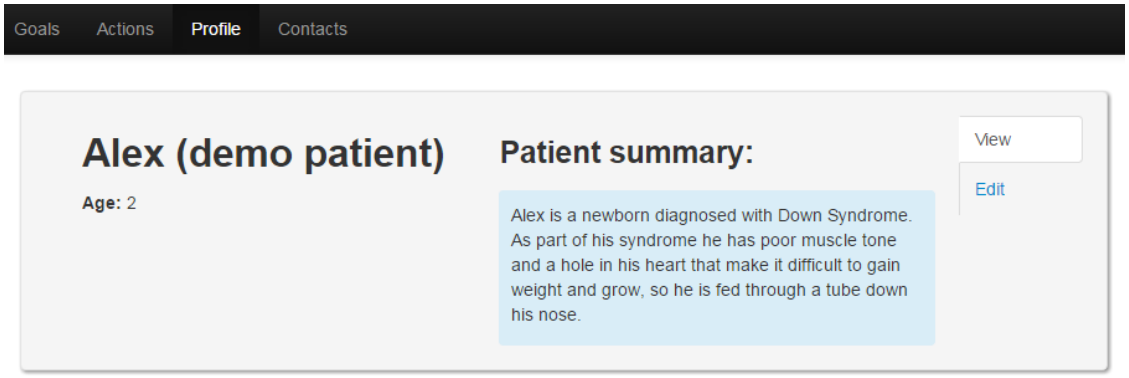

Figure 5.5: The "Profile" page. Displays general information about the patient.

system was the mother of a teenage girl with special healthcare needs (henceforth we refer to this study participant as P1). For the other family, the user was the patient herself, who was 19 at the time of the study and managed her own care (henceforth we refer to this study participant as P2).

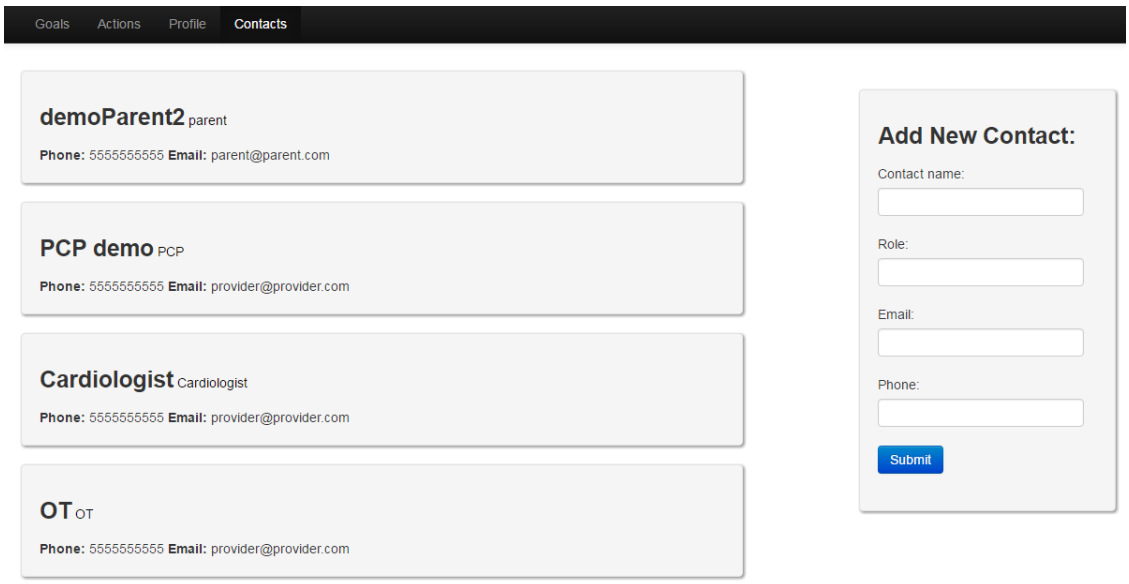

Figure 5.6: The "Contacts" page. Displays contact details for caregivers.

5.3.2 Procedure

The participants were asked to use GoalKeeper for a duration of one month. At the beginning of the study, we conducted a training session that lasted 90 minutes. The training session included two parts: training on goal setting and training on the use of GoalKeeper. The goal setting training was developed and delivered by one of our pediatrician collaborators from Stanford, who joined the training session remotely. Participants were given a fictional scenario of a complex patient and were asked to think about goals for this patient, actions that might be needed to achieve these goals and metrics they might want to monitor to assess progress toward these goals. Throughout this exercise, the pediatrician discussed with the participants ways in which they can go about setting goals and identifying actions and relevant metrics. Then, the parent participant was asked to think of goals for the child, and the patient participant was asked to think of goals for herself. The goal setting training materials are provided in Appendix A.1.

Following the goal setting training, participants were guided through the use of GoalKeeper.

To ensure participants were able to use the system, they were asked to input the goals and actions they wrote during the goal setting training such that they left the initial training with a couple of goals and actions already set in GoalKeeper. Participants were also provided with a tutorial to which they could refer later. Lastly, each participant was provided with a tablet as compensation for their participation. The purpose of providing a tablet was that participants could show the information in GoalKeeper to providers at clinic visits.

After the training, participants used GoalKeeper for a month. They were instructed to record status updates at least a couple of times a week. After two weeks, participants were sent a mid-study survey. At the end of the study, we conducted exit interviews with participants (see survey and interview questions in Appendix A.2).

5.3.3 Findings

System use: Table 5.1 provides a quantitative summary of participants' use of GoalKeeper. P1 used GoalKeeper consistently throughout the study. She created the first three goals at the onset of the study and added the fourth goal a couple of weeks into the study after she felt one of the other goals (trial of medication) was completed successfully. The first goal was to complete a trial for a new medication. Actions for this goal included scheduling appointments and picking up the medication. Status updates for this goal reported whether her daughter took the medication or not. Other goals were related to the use of medical equipment for improving physical movement (wearing Dynamic Movement Orthosis and a splint). Actions for these goals were concerned with handling the equipment and scheduling appointments, and status updates recorded the use of the equipment. The fourth goal was added in the last week of the study, upon completion of the medication trial (the first goal). The status updates for this goal reported the amount of time the patient exercised during the day.

Typically, notes were added to each status update. In some cases, these comments added

	Goal	# Actions	# Status updates
P1	(1) Trial of medication	6	10
	(2) Wear DMO	2	15
	(3) Wear splint	2	10
	(4) Daily exercise	0	3
P2	(1) Take medication	1	4
	(2) Use blue light	0	3
	(3) Class attendance	2	2
	(4) Mood	0	2

Table 5.1: Participants' use of GoalKeeper.

more details about the status that was recorded. For example, medication dosage was recorded for status updates about medication trial and the duration of wearing the medical device was reported for goals (2) and (3). In other cases, notes reflected positive or negative sentiment about progress towards the goal. For instance "we are on a roll" or "UGH...we can't seem to make this work". In several instances the notes described the obstacles the family faced, for example "until we get a schedule, can't seem to initiate wearing".

P2 used the system in the first two weeks but later stopped using it when she got sick (we discuss this later in this section). She created four goals at the beginning of the study. For the first two goals, taking medication and using a blue light, status updates were binary (whether or not the drug was taken/blue light was used). The third goal was to increase her class attendance and for this goal P2 recorded the number of classes she attended and commented on the ones she skipped. Lastly, P2 tracked her mood and recorded numeric values describing her level of happiness. P2 typically did not add notes when recording status updates. The few actions P2 inputted into the system were related to scheduling appointments.

In the mid-survey questionnaire, both participants rated the system as easy to use. They both rated the goal setting feature (5 and 6 out of 7), action tracking feature (6 and 4 out of 7) and status update feature (5 and 7 out of 7) as useful. Neither participant made use of the Contacts and Profile pages.

Goal setting process: P1 set the goals for her child by herself. She decided on the

goals by thinking about the immediate needs of her child: "When I thought about the goals, I really thought about what kind of some of the immediate needs were and kind of worked from there, like things that I knew were kind of front and center for us. I mean there are other longer range goals [...] I didn't put in there." One of the goals arose in a discussion with a provider: "we were struggling [...] and then we had a visit with the provider and she kind of set the course, and so then we had a goal. That really cemented the goal like okay now we have, I think we have the tools we need to make this goal happen." P2 set the goals together with her therpist; she first came up with the goals for herself and then discussed them with the therapist.

P1 commented on the goal setting process as an organizing mechanism for managing care: "it made me more aware that goal setting requires organization of activities [...] you start to get a little disorganized based on the scale of things that have to be done, and so the organization of this in and of itself is just helpful in that now I don't have to store it."

Working toward goals and tracking progress: P1 found the recording of status updates useful for quantifying more accurately progress. Moreover, tracking progress often prompted her to try to think of new actions to take: "[...] as problems arose, I thought about the goals and it was like 'Oh this is not working, we are having a problem here'; now I have an action that I can take related to that. I've got to call PT or I've got to follow up with someone.". Tracking the status updates also helped identify reasons for failing at working toward a goal, which sometimes led to new strategies: "[...] when I can really start to see patterns like, oh on certain days she doesn't want to do it, is their school day too busy, what are some creative ways that I can kind of try to inspire her to do something?" Reviewing the status updates also helped P1 identify when a goal was achieved "Yeah we are done, we were trialing it, we struggled when we started and then now we are successful."

Both participants described how viewing the progress, and in particular the lack of progress, toward a goal affected them emotionally. P1 commented that completing actions

"felt like good, I got something done. So that part of it I liked." Similarly, when P2 saw she did well with respect to her goal to increase class attendance, it encouraged her to continue: "[...] recording how often I was going to class made it easier for me to go to class, because I didn't want to break the chain. I had to continue going so that the graph would still look good [...] it encouraged me to do it because I saw how good I was doing and I wanted to keep doing good." On the other hand, at a later time, the negative emotional effect when not making progress caused P2 to stop using GoalKeeper: "I did start getting sick and I stopped doing everything and all of my responsibilities, so then the graph went down and then I got kind of upset so I stopped using it for a little bit [...] And then I stopped going to classes. And then I didn't want to like put it in."

Maintaining engagement with GoalKeeper was sometimes challenging to both participants. They commented that reminders would be helpful to ensure that they record information consistently. P1 commented that although entering information daily sometimes felt like too much "there is something to be said for the habit of entering it daily, because then once you habituate it then it becomes like 'Oh I'm willing to do it frequently'; it is easy to slip off and not do it at all."

The version of GoalKeeper used in the study supported only numerical status updates. Both participants used workarounds for status updates of different nature. For example, they used '1' and '0' to describe whether or not a medicine was taken.

Sharing information with providers: Both participants commented about how they shared or envisioned sharing information from GoalKeeper with care providers. P2 had an appointment with her therapist during the study commented that while she did not show him GoalKeeper, it helped her report about her progress. When asked whether she would like her therapist to be able to view her goals and progress, she said "I think during our appointments it definitely would have been helpful. But I don't know how I would feel about them seeing it 24/7."

P1 did not have meetings with providers during the study, but thought that GoalKeeper could also be useful for communication with providers: "I see the big value is that then it can also be a connection to the provider. If that– if it's a shared space". However, she commented that the use of the system with each provider would need to be defined within the caregiver-family relationship because caregivers tend to be different in the information they are interested in, and families differ in the level of detail they provide as well. Specifically, she was concerned with respect to the patient summary feature that "people may just put in unlimited amount of information in there. And then it is too much and then the doctors are going to be like 'I can't, I'm not going to use this because there is more information in here than I need'."

5.4 Discussion

The initial pilot study of GoalKeeper suggests potential benefits of a care plan management system, as well as challenges to making such a system useful and usable. We next summarize the key results of the study and its limitations, and suggest possible directions for future designs of GoalKeeper.

Key findings and design implications. According to study participants, GoalKeeper helped them better organize their activities. More importantly, the need to explicitly describe goals and track them made participants more aware of successes and failures, which in some occasions led them to take action (e.g., adding a new goal when an existing goal was achieved). This outcome is particularly encouraging, as patient (if the patient is old enough) and family activation have been shown to have positive impact on health outcomes [Larson and Reid, 2010]. Future versions of GoalKeeper could further support family engagement by automatically drawing users' attention to goals that may require attention, either due to lack of progress or decline, or due to consistent positive improvement. In addition, GoalKeeper

could incorporate new features to support families' agenda preparation for clinic visits. For example, it could summarize the status of different goals and enable families to describe their main concerns and questions prior to their meetings with care providers.

While participants saw value in tracking the status of goals and actions, they found the need to input information on a daily basis burdensome. This problem might be alleviated to some extent with the development of a mobile application version of GoalKeeper, eliminating the need to open a tablet or a computer just to input a status update. Future designs could also make it easier for users to remember to record status updates by enabling them to set reminders or by proactively prompting them.

Applications for supporting behavior change in areas such as exercise and nutrition often encourage users by highlighting their successes and reinforcing positive behaviors [Consolvo et al., 2008]. In complex care settings, however, the progress or lack of progress toward goals may be entirely out of the hands of the family in many situations (e.g., a drug not working). Therefore, highlighting achievements will not always be possible. Our initial study revealed that coping with negative outcomes can be difficult. Both participants mentioned the negative emotional effects of such situations, and one of them even stopped using GoalKeeper for several days to avoid seeing her failure in achieving one of her goals. This presents an important design challenge of maintaining engagement during hardships, and supporting families in identifying situations in which the plan does not seem to work and in initiating a discussion with relevant providers to modify the plan accordingly.

Limitations. The pilot study only involved two users and additional studies are required to draw a wider range of design implications for future designs of GoalKeeper. In addition, a key question not addressed in our pilot study is whether and to what extent providers will engage with GoalKeeper. Engaging providers may be particularly difficult as they are already overwhelmed by their interactions with existing electronic medical records systems. Addressing this design challenge requires further studies of providers' needs and workflows.

5.5 Related Work

In this section, we review prior work on patient self-tracking and sharing of patient collected data with care providers. Background about the use of care plans in complex care is described in Section 2.1.

Prior work on personal health information management (PHIM) has explored how patients manage their medical records and proposed the design of systems to improve people's ability to manage their health information [Pratt et al., 2006, Piras and Zanutto, 2010]. A study of cancer patients found that they often need to integrate personal and health information from different sources such as prescriptions, medical notes and emails. This fragmentation of information makes it hard for patients to gain a better understanding of the overall picture and understand what this information means for their health-related decisions [Pratt et al., 2006].

Klasnja et al. [2010b] developed HealthWeaver, a mobile application for helping cancer patients manage information about their care. This system enabled patients to track their symptoms and prepare for clinic visits. A study of HealthWeaver [Patel et al., 2012] suggested several design implications for tracking tools, including providing patients with guidance on what metrics to track, enabling customization for different patient needs, supporting reflection and communication with clinicians by visualizing data in meaningful ways and giving patients ownership over data tracking. Our design of GoalKeeper follows these guidelines. In addition, it aims to support families in contextualizing actions and health status with respect to the patient's care goals. Mamykina et al. [2008] developed a health monitoring system for diabetes patients. They showed that showed that self-tracking with their system led patients to adopt an Internal Locus of Control, which has been shown to increase engagement and improve health outcomes.

Chung et al. [2016] studied patients' perspectives and collaboration practices around

the use of patient-generated data in the care for patients with irritable bowel syndrome or obesity. They found that patients had varied expectations about the use of generated date, such as supporting diagnosis and personalization of their treatment, improving their self-awareness and accountability and getting recognition from their providers for their efforts. They further examined several case studies of interactions between patients and providers in which patient-generated data was shared, and suggest ways in which patient-generated data can help patients and providers develop *boundary negotiating artifacts* [Lee, 2007]. These artifacts can help push the boundaries between patients and providers and support their collaboration. We intend GoalKeeper to also support collaboration between patients and providers, as well as between different care providers, by helping the team to establish and track a mutually agreed-upon care plan.

Chapter 6

Interactive Teaching Strategies for Agent Training

When agents are face the task of learning how to act in a new environment, they can benefit from the input of more experienced agents and humans. This chapter presents interactive agent training heuristics in the context of a student-teacher reinforcement learning framework. In this framework, an experienced "teacher" agent helps accelerate the "student" agent's learning by providing advice on which action to take next [Clouse, 1996, Torrey and Taylor, 2013]. The student updates its policy based on reward signals from the environment as in typical reinforcement learning, but its exploration is guided by the teacher's advice.

Prior work has considered two modes of advice-giving in this framework: student-initiated [Clouse, 1996] and teacher-initiated [Torrey and Taylor, 2013]. Torrey and Taylor [2013] considered a setting with a limited advice budget and developed heuristics that guide the teacher's choice of advising opportunities. They demonstrated significant learning gains when using these heuristics in empirical studies. While the amount of advice that the teacher can provide was limited, their formulation assumed that the student's current state is always communicated to the teacher, and that the teacher continuously monitors the student's

decisions until the advice budget runs out. These assumptions have significant drawbacks. For human teachers, constantly paying attention diminishes the value of automation, imposes cognitive costs [Miller et al., 2015] and can be simply unrealistic. Even if the teacher is a computer agent, transmitting the student's every state to the teacher can have a prohibitive communication cost.

We investigated interactive student-teacher training, in which the student *and* the teacher jointly decide when advice should be given. In these *jointly-initiated* teaching strategies, the student determines whether to ask for the teacher's attention, and the teacher, if asked to pay attention to the student's state, decides whether to use this opportunity to give advice, given a limited advice budget. We begin by comparing the teacher-initiated and student-initiated approaches experimentally, showing that heuristics for teacher-initiated training are more effective at improving the student agent's policy than student-initiated ones, but they require more teacher attention. Then we demonstrate that the jointly-initiated teaching strategies can reduce the amount of attention required of the teacher compared to teacher-initiated strategies, while maintaining similar learning gains. Thus, our approach integrates the teacher-initiated and student-initiated approaches, alleviating their disadvantages.

Collaborative approaches for assisting agents are particularly important for semi-autonomous agents [Zilberstein, 2015] (e.g., self-driving cars), as such agents will have long-term interactions with people and will have opportunities to continuously improve their policies based on these interactions. Therefore, in addition to comparing the effectiveness of different interactive training strategies, we investigate the effect on learning performance of factors that may vary across agent-human settings. In particular, our empirical evaluations analyze the effect of the information communicated to the teacher and the quality of the initial policy of the student on teaching outcomes.

6.1 Student-Teacher Reinforcement Learning

The student-teacher framework [Clouse, 1996] includes two agents: a student and a teacher. We assume that the teacher has already established a fixed policy for acting in the environment, denoted $\pi_{teacher}$, whereas the student uses a reinforcement learning algorithm to learn its policy, denoted $\pi_{student}$. At any state s, the teacher can give advice to the student by sharing $\pi_{teacher}(s)$. This formulation requires that the teacher and the student share the same action space but does not assume they share the same state representation. When the student receives advice from the teacher, it takes the suggested action. That action is then treated as any other action chosen by the student during the learning period, and Q-values are updated using the same learning algorithm.

Similarly to Torrey and Taylor [2013], we specify a limited advice budget for the teacher, but in contrast, we do not assume constant monitoring of the student by the teacher. Rather than specifying an attention budget, we consider the attention required of the teacher as an additional metric by which we evaluate the different teaching strategies. We consider two different metrics for attention: (1) the number of states in which the teacher had to assess the student's state (to decide whether to give advice), and (2) the overall duration of the teaching period (i.e., the last time step in which the teacher had to assess the student's state).

Instead of fixing an attention budget, we choose to analyze the amount and duration required by training strategies, because considerations about attention vary among different settings. For example, if a person is helping an autonomous car to improve its policy, the overall duration of the teaching period does not matter because the person is always present when the car drives. However, the person might not pay attention to the road at all times and therefore there is a cost associated with monitoring the car's actions to decide whether to intervene. Moreover, if teaching in this setting requires the human to take control over the car, then providing advice incurs an additional cost beyond monitoring the agent's behavior

(i.e., deciding whether to intervene requires less effort than actually intervening). Thus, in this setting, we would like to minimize the number of states in which we require the teacher's attention as well as the number of times the teacher is required to give advice. In contrast, if an expert is brought to a lab to help train a robot, teaching is done during a dedicated time period in which the teacher watches a robot student. Here, minimizing the overall duration of the teaching period will be more important than minimizing the number of states in which attention is required.

6.1.1 Teacher-Initiated Advising

Torrey and Taylor [2013] proposed several heuristics for the teacher to decide when to give advice, using the notion of state *importance*. Intuitively, a state is considered important if taking a wrong action in that state can lead to a significant decrease in future rewards, as determined by the teacher's Q-values. Formally, the importance of a state, denoted $I(s)$, is defined as:

$$I(s) = \max_a Q^{teacher}_{(s,a)} - \min_a Q^{teacher}_{(s,a)} \tag{6.1}$$

Three variations of this heuristic were suggested: (1) ***Advise Important***: giving advice when $I(s) > t_{ti}$ (where t_{ti} is a predetermined threshold); (2) ***Correct Important***: giving advice if $I(s) > t_{ti}$ and $\pi_{student}(s) \neq \pi_{teacher}(s)$. This heuristic assumes that the teacher has access to the student's chosen action for the current state; (3) ***Predictive Advising***: giving advice if $I(s) > t_{ti}$ and the teacher predicts that the student will take a sub-optimal action. This approach assumes that the teacher does not know the student's intended action and instead develops a predictive model of the students' actions over time. We do not include *Predictive Advising* in our study to avoid the assumption that a person would develop a predictive model of the students' actions. Moreover, even with agent teachers, the ability to predict an action will greatly depend on the size of the action space.

We also evaluate the ***Early Advising*** baseline heuristics used by Torrey & Taylor. Using the *Early Advising* heuristic, the teacher gives advice in all states until the entire advice budget is spent. Similarly, with ***Early Correcting***, the teacher advises the student in any state in which $\pi_{student}(s) \neq \pi_{teacher}(s)$ until exhausting the advice budget (as *Correct Important*, this heuristic assumes that the student's intended action is communicated to the teacher).

6.1.2 Student-Initiated Advising

We consider several heuristics for the student agent to determine when to ask the teacher for advice. Similarly to *correct important*, the ***Ask Important*** heuristic uses the notion of state-importance to decide whether the student should ask for advice. It uses the *student's* Q-values when computing Equation 6.1 and asks for advice when $I(s) > t_{si}$, where t_{si} is a threshold set for the student agent.

The ***Ask Uncertain*** heuristic [Clouse, 1996] also considers Q-values differences (Equation 6.1) to decide whether to ask for advice, but differs from *Ask Important* in that it asks for advice when the difference is *smaller* than a given threshold t_{unc} (Equation 6.2). Intuitively, low Q-value difference signals that the student is uncertain about which action to take. This heuristic asks for advice when:

$$\max_a Q^{student}_{(s,a)} - \min_a Q^{student}_{(s,a)} < t_{unc}, \tag{6.2}$$

where t_{unc} is the given student's threshold for uncertainty.

Chernova & Veloso [Chernova and Veloso, 2007] used the distance from a state to its previously visited nearest-neighbor state as one measure of confidence that is based on the agent's familiarity with the state. We implement this approach in the ***Ask Unfamiliar*** heuristic. In our settings states are described by a feature vector and we use Euclidean

distance between feature vectors to determine the nearest neighbor. The student then asks for advice when:

$$distance(s, NN(s)) > t_{unf}, \tag{6.3}$$

where $NN(s)$ is nearest neighbor of state s.

6.1.3 Jointly-Initiated Advising

The student-initiated and teacher-initiated advising approaches both have shortcomings. The teacher-initiated approach requires the teacher to always pay attention. On the other hand, the advising decisions of the student are likely to be weak since they are guided by the student's noisy Q-value estimates. We design *jointly-initiated* advising approaches to address these shortcomings by having the right division of tasks between the teacher and the student. These approaches do not require the teacher to pay continuous attention while still utilizing the more informed signal of the teacher about whether advice is beneficial in a given state.

In *jointly-initiated* advising, the student decides whether to ask for the teacher's *attention* based on one of the student-initiated approaches for asking for advice. Then, the teacher decides whether to provide advice based on one of the teacher-initiated advising approaches. We denote a jointly-initiated heuristic by [X–Y], where X is a student heuristic for asking the teacher's attention and Y is a teacher heuristic for determining whether to give advice in the current state. For instance, ***[Ask Important–Correct Important]*** means that the *student* asks for the teacher's attention when $I(s)_{student} > t_{si}$. The teacher will then assess the state and will give advice if $I(s)_{teacher} > t_{ti}$.

Once the teacher decides to give advice, it will continue monitoring the student's actions until advice is no longer needed, and will later resume monitoring only when the student asks

for the teacher's attention next[1]. The motivation for this approach is that once the teacher is already paying attention, it will be better able to judge whether additional advice is required in consequent states, until the student takes the right course of action. In addition, it requires less context-switching of the teacher.

6.2 Empirical Evaluation

Our experiments have five objectives: (1) Comparing student-initiated and teacher-initiated strategies: we assess the relative strengths and weaknesses of the existing approaches; (2) Evaluating the proposed jointly-initiated approaches: we compare the performance of the joint heuristics to that of the best-performing prior heuristics, as determined by the first experiment; (3) Exploring the effect of the student's initial policy quality on performance: in real-world settings, autonomous agents will likely start with some pre-programmed basic policy rather than learn "from scratch". Therefore, we evaluate the benefits of the teaching sessions when varying the quality of the student's initial policy. The quality of the initial policy is varied in two ways: by varying the length of the student's independent training (without access to teacher advice) prior to the teaching session, and by pre-training the student in limited settings that do not include some important features of the game, so that the student cannot learn certain skills; (4) Exploring the effect of sharing the student's intended action with the teacher while sharing the student's intended action can reduce the use of the teacher's advice budget, sharing the student's action might be infeasible in some domains, and also incurs additional communication costs. Therefore, we explore the extent to which sharing the intended student action benefits learning; (5) Assessing the sensitivity of the heuristics to the chosen thresholds: each of the heuristics used for deciding when to

[1]We evaluated the student-initiated approaches using this continued monitoring, but it did not lead to significant differences.

Figure 6.1: The Pac-Man Game.

advice uses a pre-defined threshold. We evaluate the robustness of these heuristics to the selection of these threshold values.

6.2.1 Experimental Setup

We used the Pac-Man vs. Ghosts League competition [Rohlfshagen and Lucas, 2011] as our experimental domain. Figure 6.1 shows the game maze used in our experiments. This game configuration includes two types of food pellets: regular pellets (small dots) are worth 10 points each and power pellets (larger dots) are worth 50 points each. In addition, after eating a power pellet, ghosts become edible for a limited time period. Pac-Man receives 200 points for each eaten ghost. A game episode ends when Pac-Man is eaten by a ghost, or after 2000 time steps. Ghosts chase Pac-Man with 80% probability and otherwise move randomly. In each state, Pac-Man has at most four moves (right, left, up or down).

Due to the large size of the state space, we use a high-level feature representation for state-action pairs. Specifically, we use the 7-feature representation from Torrey and Taylor

[2013] implementation. Q-values are defined as a weighted function of the feature values $f_i(s, a)$:

$$Q(s, a) = \omega_0 + \sum_i \omega_i \cdot f_i(s, a) \tag{6.4}$$

The student agent employed the Sarsa(λ) algorithm to learn the weights in Equation 6.4. We used the same parameter configuration as Torrey and Taylor [2013]: $\epsilon = 0.05, \alpha = 0.001, \gamma = 0.999, \lambda = 0.9$. The teacher agent was trained with the same learning configuration until its performance converged. Table 1 summarizes the heuristics for teacher-initiated and student-initiated advising strategies studied in our experiments, together with the thresholds we used for each of them. The thresholds were determined empirically.

6.2.2 Evaluation Metrics

We evaluate the student's *learning rate* by assessing the student's performance at different time points during training. Specifically, in each trial, we paused training after every 100 game episodes and evaluated the student's policy at that time point by averaging 30 evaluation episodes (in which the student uses its current policy without exploring or updating its Q-values). Because trials have high variance depending on the student's eploration, we generate a learning curve by aggregating 30 separate trials. For example, Figure 6.2 (left) shows the learning curves comparing the performance of teacher-initiated and student-initiated approaches. The x-axis represents training episodes, and y-axis values show the average episode reward at that point in the training session.

For the teacher, *cumulative attention* is evaluated by averaging the number of states in which the teacher was asked to monitor the student. We assume that the teacher completely stops monitoring once the advice budget is fully used. Cumulative attention curves are generated by averaging the total number of states in which the teacher's attention was required after every 100 game episodes, averaging these values over 30 trials. The left plot

of Figure 6.2 shows a cumulative attention curve. As in the learning curve, the x-axis corresponds to training episodes. The y-axis values show the number of states in which the teacher's attention was required up to a given time point.

The overall *attention duration* required from the teacher is the average number of states it takes to use the entire advice budget. This metric can also be assessed by looking at the x-value of the point in which the cumulative attention (y-value) flattens in the cumulative attention curves. For example, in Figure 6.2 (right), when using the *correct important* heuristic, the overall duration of required attention is 90 episodes, while *early correcting* only requires an attention duration of 10 episodes as the advice budget gets used quickly.

To assess the statistical significance of differences in average rewards and cumulative attention, we ran paired t-tests comparing the averages after each 100 training episodes.

Heuristic	**Initiator**	**Threshold**	**Shared Action**
Early Correcting	teacher	None	Yes
Early Advising	teacher	None	No
Correct Important	teacher	200	Yes
Advise Important	teacher	200	No
Ask Important	student	50	Yes
Ask Uncertain	student	30	Yes
Ask Unfamiliar	student	Avg. distance to nearest neighbor	Yes

Table 6.1: Student-initiated and teacher-initiated heuristics.

6.2.3 Teacher Vs. Student Heuristics

Figure 6.2 (left) shows the student's learning rate when using heuristics for teacher-initiated and student-initiated advising. When using the teacher-initiated approaches, the teacher constantly monitors the state of the student until the advice budget runs out. When student-initiated advising approaches are used, the teacher only monitors the student when it is asked to advise. In all cases, the student's intended action is available to the teacher when giving advice.

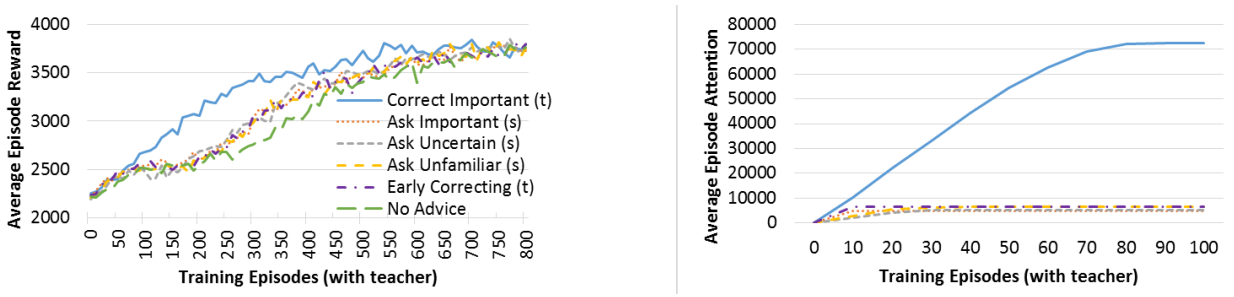

Figure 6.2: Average reward (left) and average attention (right) obtained by student-initiated and teacher-initiated approaches. The student was trained for 100 episodes prior to teaching, and actions were shared with teacher.

Substantially and significantly higher learning gains were obtained when using the teacher-initiated *Correct Important* heuristic compared to all other heuristics ($p < 10^{-16}$). This can be seen in Figure 6.2 (left). For example, after 200 training episodes with the teacher, an average reward of 3055.64 is obtained when using *Correct Important*, compared to an average reward of 2688.03 when using the next best heuristic. All other heuristics led to higher learning rates compared to the *no advice* (green dashed line) condition ($p < 10^{-10}$). There were no statistically significant differences between any other pairs of advising strategies.

The higher learning gains obtained when using teacher-initiated advising are expected; the teacher has more knowledge than the student about the domain and has a good policy for making decisions in it, which allows it to choose effective teaching opportunities. Consider the game state shown in Figure 6.1 as an illustrative example. Intuitively, this is an important state: if Pac-Man (the student) makes a wrong move, it might be eaten by a ghost; if, however, it proceeds towards the power pellet, it will have an opportunity to earn a high reward for eating a ghost. The teacher, which already knows the environment, can identify that this state is important based on its Q-values, while the student might not yet have enough information to come to this conclusion.

While the *Correct Important* heuristic results in the highest learning gains, it requires significantly more teacher attention than the other approaches. This can be seen in Figure 6.2

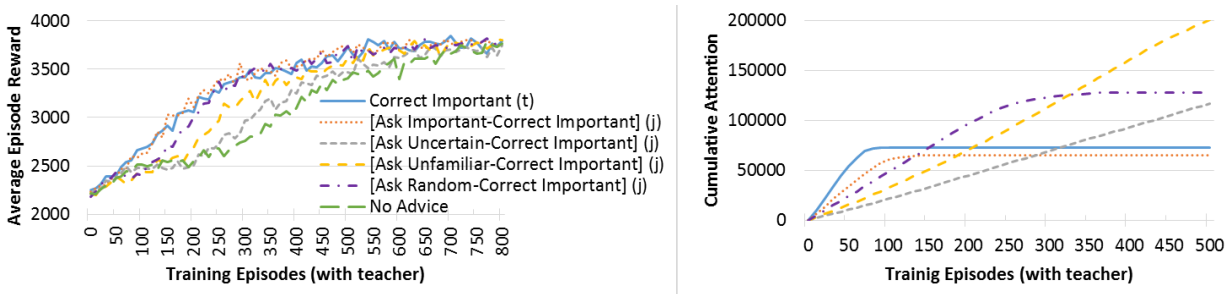

Figure 6.3: Average reward (left) and cumulative attention (right) obtained by jointly-initiated and teacher-initiated advising. The student was trained for 100 episodes prior to teaching, and actions were shared with teacher.

(right), which shows the average cumulative attention required of the teacher; i.e., the total number of states in which teacher's attention was required up to a given episode. Teacher attention is required in significantly more states when using *Correct Important* compared to all other heuristics (72382.06 states compared to only 6358.86 states for *Ask Unfamiliar*, which is the most attention-demanding student-initiated heuristic, $p < 0.0001$). In addition, the overall duration of teacher's attention (i.e., number of episodes until the advice budget is fully used) is larger for *Correct Important* (90 episodes compared to less than 40 episodes required when using any of the other heuristics).

6.2.4 Jointly-Initiated Teaching Strategies

The results reported so far show that the teacher-initiated advising strategies outperform the student-initiated ones in terms of learning gains, but require more attention. In this subsection, we present results from an evaluation of the jointly-initiated teaching strategies, which aim to reduce the attention required from the teacher while maintaining the benefits of student-initiated advising. We thus compare their performance with that of the top performing teacher-initiated teaching strategy, *Correct Important*.

As Figure 6.3 (left) shows, when using the heuristic *[Ask Important–Correct Important]*, the student obtains similar rewards to those obtained when using *Correct Important*. The

rewards at any given time point were on average slightly higher when using *[Ask Important–Correct Important]*, but while this difference was statistically significant ($p = 0.008$), it was not substantial (average difference of 18.5 points). Figure 6.3 (right) shows that the *[Ask Important–Correct Important]* heuristic required the teacher's attention in fewer states (64,711.47 states compared to 72,382.07 states). This difference was statistically significant ($p < 10^{-5}$), and substantial. However, the overall duration of required teacher's attention when using *[Ask Important–Correct Important]* is 140 episodes (indicated by the x value corresponding to the maximal total attention), compared to 50 episodes when using the *Correct Important* heuristic. That is, while the jointly-initiated teaching strategy requires the teacher's attention in fewer states, the duration of the training session, and thus teacher's needed attention span, is longer.

To ensure that the performance is achieved as a result of the student's choice of states in which to ask for advice, we also evaluate a random baseline where the student asks for the teacher's attention with 0.5 probability (the average rate of asking for advice by the *Ask Important* heuristic until the advice budget runs out). As shown in Figure 6.3, this random baseline (*Ask Random–Correct Important]*, dashed purple) does not perform as well as *[Ask Important–Correct Important]*. Moreover, it requires significantly more cumulative teacher attention, as well as a longer teaching period (both attention and learning gains differences were statistically significant, $p < 10^{-5}$). This shows that while the student's perception of importance is not as accurate as that of the teacher, it is still useful for identifying advising opportunities.

The strength of the *[Ask Important–Correct Important]* heuristic is its recall for important states. While the student heuristic *Ask Important* has many false positives when trying to identify important states due to the students' inaccurate Q-values, combining it with the teacher's *Correct Important* heuristic, which assesses whether the state is truly important, mitigates this weakness.

The other jointly-initiated teaching strategies, *[Ask Uncertain–Correct Important]* and *[Ask Unfamiliar–Correct Important]*, lead to some improvement in learning rate compared to the *No Advice* baseline, but perform significantly worse than *Ask Random* and require more cumulative attention, because they rarely capture important states. That is, they suffer from a high false negative rate when trying to identify important states, and therefore when the teacher uses *Correct Important* in combination with these approaches, it typically decides not to give advice (as the state is not important). This is evident by the long duration it takes until the advice budget runs out when using these heuristics (Figure 6.3).

[Ask Uncertain–Correct Important] suffers from a higher false negative rate because in its essence, *Ask Uncertain* captures states with a small Q-value range rather than those with a high one. While in some of these states the student might be uncertain of its actions, it might also mean that none of the actions will lead to significantly decreased performance. *[Ask Unfamiliar–Correct Important]* also suffers from possibly missing important states and using the advice when its impact is smaller, because unfamiliar states might not be important ones. In addition, appropriately identifying unfamiliar states likely requires more sophisticated domain-dependent similarity methods.

6.2.5 The Effect of Student's Initial Policy

The quality of the student's initial policy may affect the effectiveness of different advising strategies. To gain insights into this relationship, we experimented with student agents that differ in the quality of their initial policies.

We observe similar trends and relative performance of the different teaching strategies when varying the length of the student's independent training prior to the teaching session. As the quality of the initial policy of the student improves (i.e., the student's initial policy is based on more independent learning episodes, in the same game settings), the performance of the jointly-initiated and student-initiated teaching strategies that are based on state

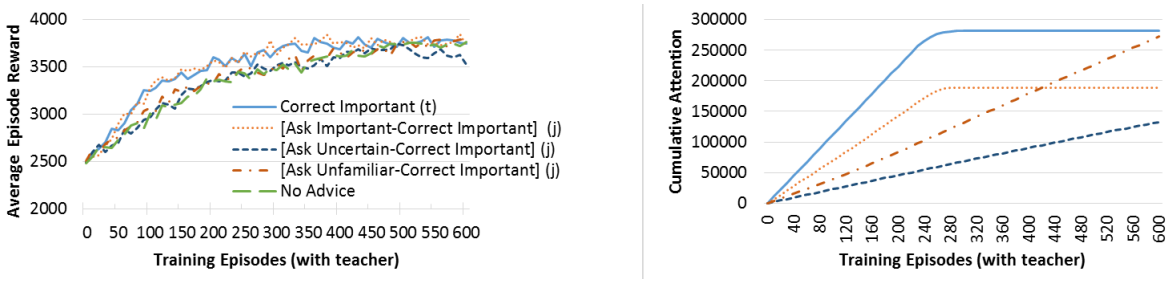

Figure 6.4: Average reward (left) and cumulative attention (right) obtained by jointly-initiated and teacher-initiated advising. The student was trained for 300 episodes prior to teaching.

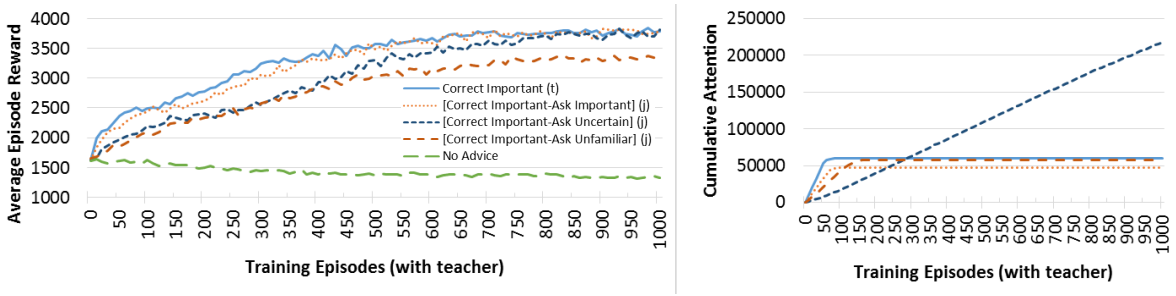

Figure 6.5: Average reward (left) and cumulative attention (right) obtained by jointly-initiated and teacher-initiated advising. The student trained for 150 episodes prior to teaching in game without power pellets, and actions were shared with teacher.

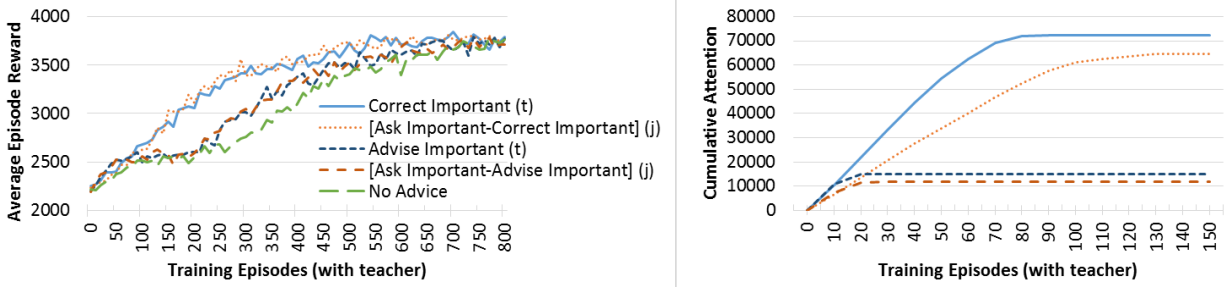

Figure 6.6: Average reward (left) and cumulative attention (right) obtained by jointly-initiated and teacher-initiated advising when the action is shared vs. when it is not shared. The student was trained for 100 episodes prior to teaching.

importance can better identify important states in which the teacher's attention is required. However, in general, the overall benefit of advising the student decreases, as there is less room for improvement of higher-quality initial policies.

Figure 6.4 shows the performance of the jointly-initiated approaches and the *Correct Important* heuristic when the student's initial policy was established after 300 episodes of individual training. While the overall trends are similar to those obtained when the initial policy was determined after only 100 episodes (Figure 6.3), the reduction in the required teacher attention when using *[Ask Important–Correct Important]* compared to that required when using *Correct Important* increases when the student starts with a better initial policy. For example, when the student was independently trained for 100 episodes, advising based on *[Ask Important–Correct Important]* required attention in 7670.6 fewer states than advising based on *Correct Important*; the difference in required attention increased to 92162.6 states if the student was trained for 300 episodes independently. In addition, when the student had longer independent training, the overall attention duration when using *[Ask Important–Correct Important]*) was only 10 episodes longer than when using *Correct Important* (compared to 50 episode gap when the independent training only lasts 100 episodes).

Teacher advice is especially beneficial when the student learns its initial policy in a limited setting that does not allow the student to explore the complete state space. Figure 6.5 shows the performance of the different teaching strategies when applied to a student that developed an initial policy in settings without power pellets. Without any advising (green bottom line), the student's policy does not improve, as it does not manage to learn about the positive rewards of eating power pellets and consequently does not learn to eat ghosts. Since the student has already established low weights for features that correspond to the possibility of eating power pellets, without guidance it is not able to learn this new skill. However, with teaching, it quickly improves its policy.

While *[Ask Important–Correct Important]* is still the best performing heuristic for jointly-

initiated advising, the *[Ask Unfamiliar–Correct Important]* heuristic does relatively better compared to settings in which the student was trained on the same game instance prior to teaching. Although it is outperformed by the *[Ask Uncertain, Correct Important]* heuristic in terms of student performance (after 400 episodes), it requires significantly less teacher attention. The relative performance improvement for the *Ask Unfamiliar* approach for getting the teacher's attention can be explained by the fact that states that involve high proximity to power pellets may appear less familiar (as they were not included in the initial independent training), and also correspond to important states.

6.2.6 Sensitivity to Thresholds

All the heuristics for choosing when to teach require specifying a threshold (the collaborative approaches require two thresholds, one for the student to decide whether to ask for attention, and one for the teacher to decide whether to give advice). To test the sensitivity of these heuristics to the specific choice of thresholds, we run experiments varying the their values.

Figure 6.7 shows the learning rate of the student when the *Correct Important* (teacher-based) heuristic is used with varied values. Reducing the advice threshold from 200 (the value used in the experiments reported in previous sections) results in lower learning rate (lighter blue dashed lines), as the advice budget is spent on less important states. However, this also reduces the amount of teacher attention, as the advice budget is used up after fewer episodes. In contrast, increasing the threshold (darker blue dashed lines) results in similar and sometimes higher learning rate, but at the cost of significantly more attention (shown on the right figure).

The collaborative heuristics require specifying two thresholds, one for the student to decide when to ask for the teacher's attention, and one for the teacher to decide when to give advice. Figure 6.8 shows the performance of the collaborative approach when using

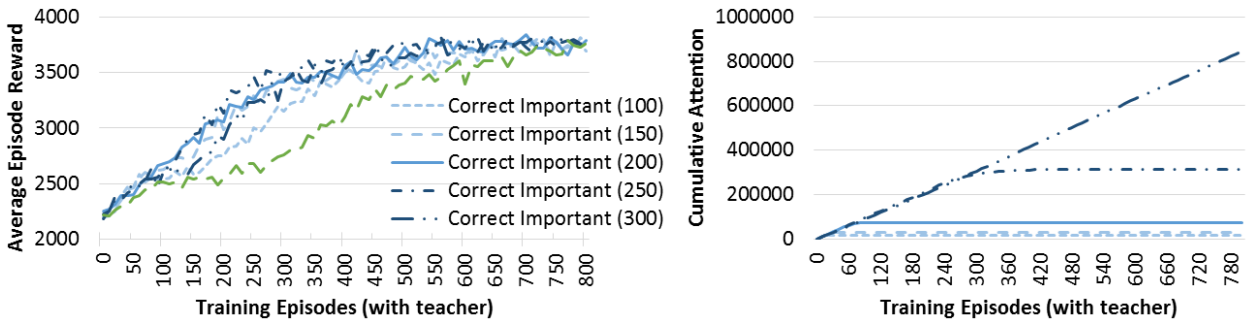

Figure 6.7: Average reward (left) and cumulative attention (right) obtained by the *Correct Important* teacher heuristic using different thresholds. Student trained for 100 episodes prior to teaching.

different thresholds for the student. Setting the threshold too high (*Ask Important(150), Correct Important (200)*) decreases the learning rate of the student, as it does not ask for help enough. Generally, increasing this threshold results increased required overall duration of the teacher's attention, as the student asks for help less frequently. Changing the teacher's threshold for giving advice in the collaborative approaches results in similar performance trends to those obtained for teacher-based *Correct Important* heuristic shown in Figure 6.7.

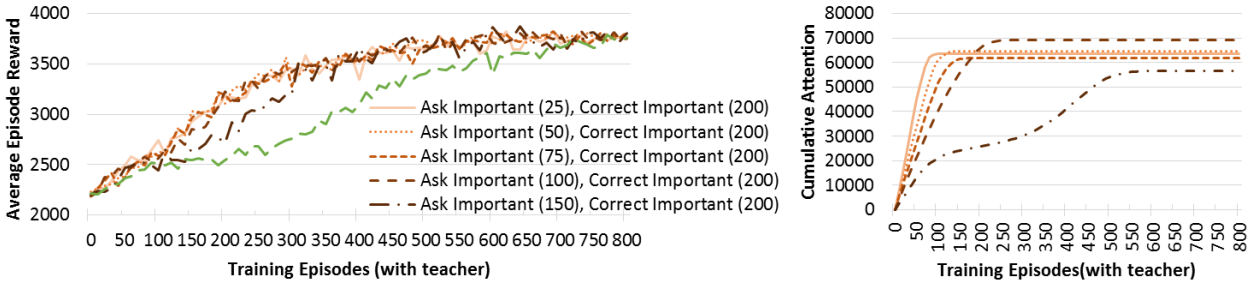

Figure 6.8: Average reward (left) and cumulative attention (right) obtained by the *Ask Important, Correct Important* teacher heuristic using different thresholds for asking the teacher's attention. Student trained for 100 episodes prior to teaching.

6.3 Related Work

Our approach builds and expands on prior work studying the student-teacher reinforcement learning framework [Clouse, 1996, Torrey and Taylor, 2013], as discussed in Section 6.1.

Chernova and Veloso [2007] proposed a confidence-based approach in which a learning agent asks for demonstrations when it is uncertain of its actions. In their approach, in contrast to the teacher-student framework, the agent only learns from the expert demonstration without receiving a signal from the environment. Judah et al. [2014] proposed a framework for active imitation learning, in which an agent can query an expert for its policy for a given state. They also assumed that the learning agent does not receive a reward signal from the environment. Furthermore, they assumed that the agent can simulate trajectories and does not query for demonstration during execution. Rosenstein et al. [2004] proposed a supervised actor-critic RL framework in which a supervisor's action is integrated with a learning agent's action. In contrast to our work, they assume a continuous action space. Griffith et al. [2013] proposed a Bayesian approach for integrating human feedback on the correctness of actions to shape an agent's policy. Rosman and Ramamoorthy [2014] developed methods for deciding when to advise an agent, but assume teachers have access to a knowledge base of common agent trajectories.

Our motivation for aiming to reduce the attention required from the teacher is rooted in prior research on human-agent interaction and human attention. These works have developed methods for detecting human attention and for incorporating it into agent decision making, taking into consideration the limited attention resources available to people as well as the costs of interruptions [Horvitz et al., 1999, 2003]. *Adjustable autonomy* approaches take into account the user's focus of attention when deciding whether to act autonomously or transfer autonomy to the user [Tambe et al., 2002, Goodrich et al., 2001]. Models of human attention are also key in developing approaches for supporting humans supervising autonomous systems [Cummings and Mitchell, 2008, Fong et al., 2002].

Chapter 7

Conclusion & Future Directions

Teamwork enables people to accomplish complex tasks by drawing on team members' different expertise and dividing the work among collaborators. However, teamwork also incurs a coordination overhead as team members need to be aware of activities of others if those activities interact with their individual actions. Lack of awareness about interacting activities can lead to coordination failures and prevent the team from achieving their goals.

This thesis presents a study of complex care teams that uncovered a set of teamwork characteristics that make coordination particularly challenging: **F**lat team structure, **L**oose-coupling of activities, **E**xtended-duration of the teamwork, **C**ontinuos revision of plans, and **S**yncopated time scales (FLECS). It identifies several opportunities for technology to support such teamwork. In particular, the thesis focuses on developing intelligent information sharing mechanisms that reason about the collaborative activity to identify the subset of information that is relevant to each of the team members. It argues that personalizing the sharing of information to different team members can limit the amount of information they need to review, and can therefore reduce coordination overhead and improve team productivity.

One major contribution of this thesis is Mutual Influence Potential Networks (MIP-Nets), a new representation that models collaborative activities without relying on an explicit

model of the team's plan. In MIP-Nets, team members and the objects they interact with are represented as nodes. The weights on edges between nodes are updated over time to reflect the extent of interaction between team members and objects, as well as the extent of interaction between different objects. The thesis defines a new algorithm, MIP-DOI, which uses MIP-Nets to quantify the degree of interest that different team members are expected to have in changes made to objects.

The thesis further describes the design and implementation of a personalized change awareness mechanism for supporting collaborative writing. In contrast with the currently prevalent approach of presenting authors with *all* of the changes made by their collaborators (e.g., using track changes features), the personalized change awareness mechanism used MIP-Nets and MIP-DOI to show authors edits made by others deemed most relevant to them.

An evaluation of the personalized change awareness mechanism demonstrates the benefits of personalizing information sharing, and thus supports the thesis of the dissertation. The personalized mechanism was compared with two baselines: a change awareness mechanism that showed *all* of the changes, and a change awareness mechanism that showed the same number of changes as the personalized mechanism, but selected the changes to share at *random*. Compared to the mechanism that presented team members with all of the changes, the *personalized mechanism resulted in significantly reduced perceived workload and significantly increased productivity of team members, without reducing the quality of the work*. Furthermore, *the changes shared by the personalized mechanism were found significantly more helpful than changes chosen at random, and the quality of the teamwork was significantly higher with the personalized than the quality of the work produced by teams when the random mechanism was used*.

In addition, the thesis presents the design and implementation of GoalKeeper, a system which aims to support and enhance the use of team-based care plans for children with complex

medical condition. It describes an initial pilot study with two families which demonstrated the potential of GoalKeeper to support families by organizing care activities around goals and more consistently tracking progress toward goals.

Lastly, the thesis introduces new advising strategies for student-teacher reinforcement learning. By involving both the student and the teacher in the decision-making process, the proposed advising strategies were able to reduce the attention required of the teacher while providing similar learning gains as the best performing teacher-initiated advising strategies.

7.1 Future Work

This thesis demonstrates that intelligent information sharing methods can enhance human teamwork by reducing coordination overhead. More broadly, the increasing use of collaboration technologies presents a great opportunity for the development of intelligent systems that go beyond simply providing an infrastructure for collaboration to support more effective and efficient teamwork. There are several exciting avenues extending the research presented in this thesis toward the development of such intelligent systems.

The MIP-Nets approach can be adapted to incorporate additional inputs. Specifically, domain knowledge can be used to initialize MIP-Nets. A first step in this direction was taken in the implementation of the personalized change awareness mechanism, which made use of the document structure. More general methods can be developed to augment MIP-Nets with domain knowledge. For example, in the healthcare domain, medical ontologies could be used in addition to observations of care providers' interactions when constructing and updating MIP-Nets.

Personalized change awareness mechanisms can be improved by designing mixed-initiative interactions which will enable users to reveal more information about their goals (e.g., authors revealing the section they plan to work on), and to provide the system with feedback about

shared information. With these additional inputs, the system could adjust the algorithm to better match the users' interests and context.

Current change awareness mechanisms, including the personalized mechanism we developed, present team members with information about changes that were made by their collaborators. An interesting direction for future research is enhancing team members' awareness of possible effects that their actions may have on other aspects of the team's work. For example, in a collaborative writing scenario, when an author edits the Results section of a paper, the system might alert her that this change might create a conflict with the text describing the results in the Introduction section. This approach suggests a shift from providing only *retrospective* change awareness, to providing *prospective* awareness of the implications of changes.

Additional studies of GoalKeeper are required to improve its design and extend its use to the entire care team. Furthermore, the information sharing methods could be incorporated into GoalKeeper, such that care providers will be not be overwhelmed with unnecessary details about others' plans, but will maintain appropriate context about the overarching team plan.

7.2 The Bigger Picture: Integrated AI and HCI Research

Developing intelligent systems that support and augment people's work requires a thoughtful combination of AI and HCI approaches that goes beyond simply integrating an existing AI method into an existing interface. The research described in this thesis began with an in-depth qualitative study of complex healthcare teams. The analysis of this study drew on an AI theory which stipulates requirements for successful teamwork to derive design implications. The formulation of a new computational information sharing problem and the assumptions

about available inputs and desired outputs were based on the qualitative study of real-world teams. Finally, the design and implementation of a personalized change awareness mechanism drew on insights from HCI work on change awareness and integrated our new AI methods to adapt the presentation of changes to different users based on the inferred roles of team members and task structure. By taking an integrated AI-HCI research approach, we were able to develop a new conceptual method of personalized information sharing and implement a system that reduced people's perceived workload and increased their productivity, without having detrimental effects on the quality of the teamwork.

More generally, AI and HCI research could benefit from drawing on theories and methods developed in both fields. Specifically, HCI studies of people's work in situ can inform AI research by identifying real-world problems that could be alleviated by incorporating AI methods and to ensure that intelligent systems work well with people. Importantly, by understanding the settings in which AI methods will operate, AI researchers could also ensure that the methods they develop make appropriate assumptions about the environment in which they will be deployed. A better understanding of people's workflows and needs can help identify assumptions made by AI algorithms that do not hold in human settings. Moreover, new interactions can be designed to help elicit inputs from people that would help the algorithms adapt to their needs.

AI research can also inform work in HCI. Theories developed in HCI are often *descriptive*, that is, they describe people's behavior, cognitive abilities, etc. In contrast, AI theories are typically *prescriptive*. They specify how computer agents should be designed to provide various guarantees on the behavior of the system. As such, they can provide insight when diagnosing failures in current settings and can inform the design of new systems. Further, new AI capabilities can be utilized to support novel interactions.

Bibliography

Joanna Abraham and Madhu C Reddy. Moving patients around: a field study of coordination between clinical and non-clinical staff in hospitals. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 225–228. ACM, 2008.

Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3): 211–230, 2003.

Sherri Adams, Eyal Cohen, Sanjay Mahant, Jeremy N Friedman, Radha MacCulloch, and David B Nicholas. Exploring the usefulness of comprehensive care plans for children with medical complexity (cmc): a qualitative study. *BMC Pediatrics*, 13(10), 2014.

Ofra Amir and Kobi Gal. Plan recognition and visualization in exploratory learning environments. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(3):16–1, 2013.

Ofra Amir, Barbara J Grosz, and Roni Stern. To share or not to share? the single agent in a team decision problem. In *Models and Paradigms for Planning under Uncertainty: a Broad Perspective*, 2014.

Ofra Amir, Barbara J. Grosz, Krzysztof Z. Gajos, Sonja M. Swenson, and Lee M. Sanders. From care plans to care coordination: Opportunities for computer support of teamwork in complex healthcare. In *CHI'15*, 2015.

Tamara Babaian, Barbara J Grosz, and Stuart M Shieber. A writer's collaborative assistant. In *Proceedings of the 7th international conference on Intelligent user interfaces*, pages 7–14. ACM, 2002.

Stinne Aaløkke Ballegaard, Thomas Riisgaard Hansen, and Morten Kyng. Healthcare in everyday life: designing healthcare services for daily life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1807–1816. ACM, 2008.

Jakob E Bardram. Temporal coordination–on time and coordination of collaborativeactivities at a surgical department. *Computer Supported Cooperative Work (CSCW)*, 9(2):157–187, 2000.

Jakob E Bardram and Thomas Riisgaard Hansen. Why the plan doesn't hold: a study of situated planning, articulation and coordination work in a surgical ward. In *Proceedings of*

the 2010 ACM conference on Computer supported cooperative work, pages 331–340. ACM, 2010.

Samuel Barrett and Peter Stone. Cooperating with unknown teammates in complex domains: A robot soccer case study of ad hoc teamwork. In *AAAI*, 2015.

Michael J Barry and Susan Edgman-Levitan. Shared decision makingâĂŤthe pinnacle of patient-centered care. *New England Journal of Medicine*, 366(9):780–781, 2012.

David W Bates. Getting in step: electronic health records and their role in care coordination. *Journal of general internal medicine*, 25(3):174–176, 2010.

Hugh Beyer and Karen Holtzblatt. Contextual design. *interactions*, 6(1):32–42, 1999.

Michael Borenstein, Larry V Hedges, Julian Higgins, and Hannah R Rothstein. Converting among effect sizes. *Introduction to meta-analysis*, pages 45–49, 2009.

Kristin L Carman, Pam Dardess, Maureen Maurer, Shoshanna Sofaer, Karen Adams, Christine Bechtel, and Jennifer Sweeney. Patient and family engagement: a framework for understanding the elements and developing interventions and policies. *Health Affairs*, 32 (2):223–231, 2013.

Henian Chen, Patricia Cohen, and Sophie Chen. How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistic-sâĂŤSimulation and Computation®*, 39(4):860–864, 2010.

Sonia Chernova and Manuela Veloso. Confidence-based policy learning from demonstration using gaussian mixture models. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 233. ACM, 2007.

Chia-Fang Chung, Kristin Dew, Allison Cole, Jasmine Zia, James Fogarty, Julie A Kientz, and Sean A Munson. Boundary negotiating artifacts in personal informatics: Patient-provider collaboration with patient-generated data. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 770–786. ACM, 2016.

Jeffery Allen Clouse. On integrating apprentice learning and reinforcement learning. 1996.

P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial intelligence*, 42(2):213–261, 1990.

Katie Coleman, Brian T Austin, Cindy Brach, and Edward H Wagner. Evidence on the chronic care model in the new millennium. *Health affairs*, 28(1):75–85, 2009.

Sunny Consolvo, Predrag Klasnja, David W McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A Landay. Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 54–63. ACM, 2008.

Mary L Cummings and Paul J Mitchell. Predicting controller capacity in supervisory control of multiple uavs. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 38(2):451–460, 2008.

Cleidson De Souza, Jon Froehlich, and Paul Dourish. Seeking the source: software source code as a social and technical artifact. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 197–206. ACM, 2005.

Cleidson R de Souza, Stephen Quirk, Erik Trainer, and David F Redmiles. Supporting collaborative software development through the visualization of socio-technical dependencies. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 147–156. ACM, 2007.

Paul Dourish and Victoria Bellotti. Awareness and coordination in shared workspaces. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pages 107–114. ACM, 1992.

Rosemary Emery-Montemerlo, Geoff Gordon, Jeff Schneider, and Sebastian Thrun. Game theoretic control for robot teams. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 1163–1169. IEEE, 2005.

Kutluhan Erol, James A Hendler, and Dana S Nau. Umcp: A sound and complete procedure for hierarchical task-network planning. In *AIPS*, volume 94, pages 249–254, 1994.

AG Fiks, S Mayne, AR Localio, EA Alessandrini, and JP Guevara. Shared decision-making and health care expenditures among children with special health care needs. *Pediatrics*, 129(1):99–107, 2012.

Geraldine Fitzpatrick and Gunnar Ellingsen. A review of 25 years of cscw research in healthcare: Contributions, challenges and future agendas. *Computer Supported Cooperative Work (CSCW)*, 22(4-6):609–665, 2013.

Terrence Fong, Charles Thorpe, and Charles Baur. Robot as partner: Vehicle teleoperation with collaborative control. In *Multi-robot systems: From swarms to intelligent automata*, pages 195–202. Springer, 2002.

Jon Froehlich and Paul Dourish. Unifying artifacts and activities in a visual tool for distributed software development teams. In *Proceedings of the 26th International Conference on Software Engineering*, pages 387–396. IEEE Computer Society, 2004.

George W Furnas. *Generalized fisheye views*, volume 17. ACM, 1986.

Yaakov Gal, Swapna Reddy, Stuart M. Shieber, Andee Rubin, and Barbara J. Grosz. Plan recognition in exploratory domains. *Artif. Intell.*, 176(1):2270–2290, January 2012. ISSN 0004-3702. doi: 10.1016/j.artint.2011.09.002. URL http://dx.doi.org/10.1016/j.artint.2011.09.002.

Sebastian Gehrmann, Lauren Urke, Ofra Amir, and Barbara J Grosz. Deploying AI methods to support collaborative writing: a preliminary investigation. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 917–922. ACM, 2015.

Claudia V Goldman and Shlomo Zilberstein. Optimizing information exchange in cooperative multi-agent systems. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 137–144. ACM, 2003.

Michael A Goodrich, Dan R Olsen, Jacob Crandall, and Thomas J Palmer. Experiments in adjustable autonomy. In *Proceedings of IJCAI Workshop on Autonomy, Delegation and Control: Interacting with Intelligent Agents*, pages 1624–1629, 2001.

Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2625–2633, 2013.

Barbara J Grosz and Luke Hunsberger. The dynamics of intention in collaborative activity. *Cognitive Systems Research*, 7(2):259–272, 2006.

B.J. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

Carl Gutwin, Kevin Schneider, David Paquette, and Reagan Penner. Supporting group awareness in distributed software development. In *International Workshop on Design, Specification, and Verification of Interactive Systems*, pages 383–397. Springer, 2004.

Jörg M Haake and Brian Wilson. Supporting collaborative writing of hyperdocuments in sepia. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pages 138–146. ACM, 1992.

Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage Publications, 2006.

Pamela Hinds and Cathleen McGrath. Structures that work: social structure, work structure and coordination ease in geographically distributed teams. In *Proceedings of the 20th conference on Computer supported cooperative work*, pages 343–352. ACM, 2006.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(65-70):1979, 1979.

Reid Holmes and Robert J Walker. Customized awareness: recommending relevant external change events. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*, pages 465–474. ACM, 2010.

Eric Horvitz, Andy Jacobs, and David Hovel. Attention-sensitive alerting. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 305–313. Morgan Kaufmann Publishers Inc., 1999.

Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. Models of attention in computing and communication: from principles to applications. *Communications of the ACM*, 46(3): 52–59, 2003.

Edwin Hutchins. *Cognition in the Wild*. MIT press, 1995.

Mikkel R Jakobsen, Roland Fernandez, Mary Czerwinski, Kori Inkpen, Olga Kulyk, and George G Robertson. Wipdash: Work item and people dashboard for software development teams. In *IFIP Conference on Human-Computer Interaction*, pages 791–804. Springer, 2009.

Kshitij Judah, Alan P Fern, Thomas G Dietterich, et al. Active lmitation learning: formal and practical reductions to iid learning. *The Journal of Machine Learning Research*, 15(1): 3925–3963, 2014.

Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M German, and Daniela Damian. The promises and perils of mining github. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, pages 92–101. ACM, 2014.

E. Kamar, Y. Gal, and B.J. Grosz. Incorporating helpful behavior into collaborative planning. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 875–882. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

Gal A. Kaminka, David V. Pynadath, and Milind Tambe. Monitoring teams by overhearing: A multi-agent plan-recognition approach. *Journal of Artificial Intelligence Research*, 2002.

Julie A Kientz, Gillian R Hayes, Gregory D Abowd, and Rebecca E Grinter. From the war room to the living room: decision support for home-based therapy teams. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 209–218. ACM, 2006.

Predrag Klasnja, Andrea Civan Hartzler, Kent T Unruh, and Wanda Pratt. Blowing in the wind: unanchored patient information work during cancer care. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 193–202. ACM, 2010a.

Predrag Klasnja, Andrea Hartzler, Christopher Powell, Giovandy Phan, and Wanda Pratt. Healthweaver mobile: Designing a mobile tool for managing personal health information during cancer care. In *AMIA Annu Symp Proc*, volume 2010, pages 392–6, 2010b.

Jun-young Kwak, Rong Yang, Zhengyu Yin, Matthew E Taylor, and Milind Tambe. Robust execution-time coordination in dec-pomdps under model uncertainty. In *Sixth Annual Workshop on Multiagent Sequential Decision Making in Uncertain Domains (MSDM-2011)*, page 39, 2011.

Eric B. Larson and Robert Reid. The patient-centered medical home movement. *JAMA: the journal of the American Medical Association*, 303(16):1644–1645, 2010.

Lucian Leape. *Order From Chaos: Accelerating Care Integration*. National Patient Safety Foundation, 2012.

Charlotte P Lee. Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work (CSCW)*, 16(3):307–339, 2007.

Hector J Levesque, Philip R Cohen, and José HT Nunes. *On acting together*. SRI International Menlo Park, CA 94025-3493, 1990.

Paul Benjamin Lowry, Aaron Curtis, and Michelle René Lowry. Building a taxonomy and nomenclature of collaborative writing to improve interdisciplinary research and practice. *Journal of Business Communication*, 41(1):66–99, 2004.

Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. Mahi: investigation of social scaffolding for reflective thinking in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 477–486. ACM, 2008.

Jeanne W. McAllister. Achieving a shared plan of care with children and youth with special health care needs, 2014.

Francisco S Melo, Matthijs TJ Spaan, and Stefan J Witwicki. Querypomdp: Pomdp-based communication in multiagent systems. In *Multi-Agent Systems*, pages 189–204. Springer, 2012.

David Miller, Annabel Sun, Mishel Johns, Hillary Ive, David Sirkin, Sudipto Aich, and Wendy Ju. Distraction becomes engagement in automated driving. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 59, pages 1676–1680. SAGE Publications, 2015.

Christine M Neuwirth, Ravinder Chandhok, David S Kaufer, Paul Erion, James Morris, and Dale Miller. Flexible diff-ing in a collaborative writing system. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pages 147–154. ACM, 1992.

Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. Dec-pomdps with delayed communication. In *The 2nd Workshop on Multi-agent Sequential Decision-Making in Uncertain Domains*, 2007.

Judith S Olson and Stephanie Teasley. Groupware in the wild: Lessons learned from a year of virtual collocation. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*, pages 419–427. ACM, 1996.

Ann S O'Malley, Joy M Grossman, Genna R Cohen, Nicole M Kemper, and Hoangmai H Pham. Are electronic medical records helpful for care coordination? experiences of physician practices. *Journal of general internal medicine*, 25(3):177–185, 2010.

Inah Omoronyia, John Ferguson, Marc Roper, and Murray Wood. Using developer activity data to enhance awareness during collaborative software development. *Computer Supported Cooperative Work (CSCW)*, 18(5-6):509–558, 2009.

Rohan Padhye, Senthil Mani, and Vibha Singhal Sinha. Needfeed: taming change notifications by modeling code relevance. In *Proceedings of the 29th ACM/IEEE international conference on Automated software engineering*, pages 665–676. ACM, 2014.

Rupa A Patel, Predrag Klasnja, Andrea Hartzler, Kenton T Unruh, and Wanda Pratt. Probing the benefits of real-time tracking during cancer care. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1340. American Medical Informatics Association, 2012.

David Pinelle and Carl Gutwin. Loose coupling and healthcare organizations: deployment strategies for groupware. *Computer Supported Cooperative Work (CSCW)*, 15(5-6):537–572, 2006.

Enrico Maria Piras and Alberto Zanutto. Prescriptions, x-rays and grocery lists. designing a personal health record to support (the invisible work of) health information management in the household. *Computer Supported Cooperative Work (CSCW)*, 19(6):585–613, 2010.

Alexander Pokahr, Lars Braubach, and Winfried Lamersdorf. Jadex: A bdi reasoning engine. In *Multi-agent programming*, pages 149–174. Springer, 2005.

Wanda Pratt, Kenton Unruh, Andrea Civan, and Meredith M Skeels. Personal health information management. *Communications of the ACM*, 49(1):51–55, 2006.

Wolfgang Prinz, Elke Hinrichs, and Irina Kireyev. Anticipative awareness in a groupware system. In *From CSCW to Web 2.0: European Developments in Collaborative Design*, pages 3–20. Springer, 2010.

David V Pynadath and Milind Tambe. The communicative multiagent team decision problem: analyzing teamwork theories and models. *Journal of Artificial Intelligence Research*, 16(1): 389–423, 2002.

Madhu C Reddy and Patricia Ruma Spence. Collaborative information seeking: A field study of a multidisciplinary patient care team. *Information processing & management*, 44(1): 242–255, 2008.

Charlie Rich, Candace L Sidner, and Neal Lesh. Collagen: Applying collaborative discourse theory to human-computer interaction.(articles). *AI magazine*, 22(4), 2001.

Philipp Rohlfshagen and Simon M Lucas. Ms pac-man versus ghost team cec 2011 competition. In *Evolutionary Computation (CEC), 2011 IEEE Congress on*, pages 70–77. IEEE, 2011.

Michael T Rosenstein, Andrew G Barto, Jennie Si, Andy Barto, Warren Powell, and Donald Wunsch. Supervised actor-critic reinforcement learning. *Handbook of Learning and Approximate Dynamic Programming*, pages 359–380, 2004.

Benjamin Rosman and Subramanian Ramamoorthy. Giving advice to agents with hidden goals. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 1959–1964. IEEE, 2014.

M. Roth, R. Simmons, and M. Veloso. What to communicate? execution-time decision in multi-agent POMDPs. *Distributed Autonomous Robotic Systems 7*, pages 177–186, 2006.

Maayan Roth, Reid Simmons, and Manuela Veloso. Reasoning about joint beliefs for execution-time communication decisions. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 786–793. ACM, 2005.

Aleksandra Sarcevic, Leysia A Palen, and Randall S Burd. Coordinating time-critical work with role-tagging. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 465–474. ACM, 2011.

Dena J Schulman-Green, Aanand D Naik, Elizabeth H Bradley, Ruth McCorkle, and Sidney T Bogardus. Goal setting as a shared decision making strategy among clinicians and their older patients. *Patient education and counseling*, 63(1):145–151, 2006.

Juliet P. Shaffer. Multiple hypothesis-testing. *Annual Review of Psychology*, 46:561–584, 1995.

Calvin Sia, Thomas F Tonniges, Elizabeth Osterhus, and Sharon Taba. History of the medical home concept. *Pediatrics*, 113(Supplement 4):1473–1478, 2004.

E. Sonenberg, G. Tidhar, E. Werner, D. Kinny, M. Ljungberg, and A. Rao. Planned team activity. *Artificial Social Systems*, 890, 1992.

Matthijs TJ Spaan, Geoffrey J Gordon, and Nikos Vlassis. Decentralized planning under uncertainty for teams of communicating agents. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 249–256. ACM, 2006.

Dan Sperber and Deirdre Wilson. Precis of relevance: Communication and cognition. *Behavioral and brain sciences*, 10(04):697–710, 1987.

Michelle Potts Steves, Emile Morse, Carl Gutwin, and Saul Greenberg. A comparison of usage evaluation and inspection methods for assessing groupware usability. In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, pages 125–134. ACM, 2001.

Christopher J Stille, Thomas J McLaughlin, William A Primack, Kathleen M Mazor, and Richard C Wasserman. Determinants and impact of generalist–specialist communication about pediatric outpatient referrals. *Pediatrics*, 118(4):1341–1349, 2006.

Bonnie Strickland, Merle McPherson, Gloria Weissman, Peter Van Dyck, Zhihuan J Huang, and Paul Newacheck. Access to the medical home: results of the national survey of children with special health care needs. *Pediatrics*, 113(Supplement 4):1485–1492, 2004.

James Tam and Saul Greenberg. A framework for asynchronous change awareness in collaborative documents and workspaces. *International Journal of Human-Computer Studies*, 64(7):583–598, 2006.

Milind Tambe. Agent architectures for flexible practical teamwork. *AAAI*, 97(1):997, 1997.

Milind Tambe, Paul Scerri, and David V Pynadath. Adjustable autonomy for the real world. *Journal of Artificial Intelligence Research*, 17(1):171–228, 2002.

Lisa Torrey and Matthew Taylor. Teaching on a budget: Agents advising agents in reinforcement learning. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1053–1060. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

Vaibhav V Unhelkar and Julie A Shah. Contact : Deciding to communicate during time-critical collaborative tasks in unknown, deterministic domains. In *AAAI*, 2016.

Frank Van Ham and Adam Perer. âĂIJsearch, show context, expand on demandâĂİ: Supporting large graph exploration with degree-of-interest. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):953–960, 2009.

Devindra Weerasooriya, Anand Rao, and Kotagiri Ramamohanarao. Design of a concurrent agent-oriented language. In *Intelligent Agents*, pages 386–401. Springer, 1995.

Feng Wu, Shlomo Zilberstein, and Xiaoping Chen. Multi-agent online planning with communication. In *ICAPS*, 2009.

Feng Wu, Shlomo Zilberstein, and Xiaoping Chen. Online planning for multi-agent systems with bounded communication. *Artificial Intelligence*, 175(2):487–511, 2011.

Ping Xuan, Victor Lesser, and Shlomo Zilberstein. Communication decisions in multi-agent cooperation: Model and experiments. In *Proceedings of the fifth international conference on Autonomous agents*, pages 616–623. ACM, 2001.

Yutaka Yamauchi, Makoto Yokozawa, Takeshi Shinohara, and Toru Ishida. Collaboration with lean media: how open-source software succeeds. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 329–338. ACM, 2000.

Chongjie Zhang and Victor Lesser. Coordinating multi-agent reinforcement learning with limited communication. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1101–1108. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

Shlomo Zilberstein. Building strong semi-autonomous systems. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 4088–4092, 2015.

Appendix A

GoalKeeper Study Materials

A.1 Goal Setting Training

In the goal setting training, participants were given the following goal setting worksheet, and were guided through the process of setting goals by a pediatrician.

Goal Setting Worksheet:

Agenda:

1. Set three goals
2. Assign providers to help you with these goals
3. Put these goals into GoalKeeper
4. Start using GoalKeeper
5. Complete survey on training session

Definitions:

- **Goal**: something that you want your child/you to be able to do for his/her/your health.
- **Action**: what you need to do to reach a goal.
- **Status Update**: tracks your progress toward a goal.
- **Provider**: any health care professional involved in your child's care such as a physician, nurse, physician assistant, occupational therapist, physical therapist.

Please use this worksheet to help you set goals for your child.

Example: Alex is a newborn diagnosed with Down Syndrome. As part of his syndrome he has poor muscle tone and a hole in his heart that make it difficult to gain weight and grow, so he is fed through a tube down his nose.

In the next week, I hope Alex can:

start taking milk through his mouth (also known as eating)

I will know Alex can eat if:

I can help Alex learn to eat by:

This is important to Alex because:

Summarize the above information into a goal for Alex:

The providers who can help Alex achieve this goal are (please include name and specialty):

Part 1: In the next week, I hope my child can:

I will know my child can do the above if:

I can help my child do the above by:

This is important to my child because:

Summarize the above information into a goal:

The providers who can help me achieve this goal are (please include name and specialty):

Part 2: In the next month, I hope my child can:

I will know my child can do the above if:

I can help my child do the above by:

This is important to my child because:

Summarize the above information into a goal:

The providers who can help me achieve this goal are (please include name and specialty):

Part 3: In the next two months, I hope my child can:

I will know my child can do the above if:

I can help my child do the above by:

This is important to my child because:

Summarize the above information into a goal:

The providers who can help me achieve this goal are (please include name and specialty):

A.2 Survey and Interview Questions

Two weeks into the study, participants were asked to answer the following questions in an online survey:

- How easy/difficult is the system to use? [1 very easy – 7 very difficult]
- Are there any aspects of the system that were especially hard for you to use?
- Were you able to provide all the information you thought important about your child's status?
- What additional information would you have liked to provide for the care team?
- On average, how much time each day did entering information in GoalKeeper require?
- How useful were each of the following features? [1 not useful at all – 7 very useful]
  - Status updates
  - Action tracking
  - Contact list
  - Patient summary (profile page)
- What features would you like added to GoalKeeper?
- Would you prefer using GoalKeeper on a mobile device?
- Do you have any other comments or suggestions?

At the end of the study, an exit interview was conducted with each of the participants. The interview included the same set of question, but with open discussion with the interviewer about their actual use of the system (based on log files). In addition, participants were asked the following questions about their experience with goal-setting and GoalKeeper:

- Did you find goal-setting to be an effective way to organize care activities?
- How did you go about setting goals?
- What did you find most challenging about setting goals?
- Did you discuss the goals with your providers?

- Did the use of goals and GoalKeeper affect your communication with providers in any way?
- Would you have liked to keep using GoalKeeper after the study ended?
- Any comments, questions or suggestions?