Human-Agent Cooperation to Support Resilience

By

Erin K. Chiou

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Industrial Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON

2016

Date of final oral examination: 08/02/2016

The dissertation is approved by the following members of the Final Oral Committee:

John D. Lee, Professor, Industrial & Systems Engineering

Pascale Carayon, Professor, Industrial & Systems Engineering

Douglas Wiegmann, Associate Professor, Industrial & Systems Engineering

Bilge Mutlu, Associate Professor, Computer Science

Nicole Werner, Assistant Professor, Industrial & Systems Engineering

Dedication

To my grandparents and parents who valued education above practically everything else, and to Bris Mueller who logged 56,160 miles in the time it took me to complete my degree, for our greater good. I learned more about cooperation from you than I did from reading several decades' worth of research.

Abstract

Advancements in automation, including increasing machine autonomy, are changing people's relationships to automation, and moving machines into our more unpredictable human world. Such developments have important implications for the resilience of joint human-automation systems. When humans and machine agents actively coordinate joint goals in complex and unpredictable environments, cooperation is required. While previous research in human-automation interaction focuses on how perceptions of performance influence reliance on or compliance with automation, this research explores social exchange factors influencing human-automation cooperation and system resilience. Specifically, it considers how different social exchange structures and different levels of an automated agent's cooperation influence joint coordination in a dynamic task environment. A microworld was developed and the study was conducted in two parts. Part 1 tests whether different levels of agent cooperation affect human cooperation, given unexpected changes in the task environment, in a negotiated exchange structure. Part 2 also tests levels of agent cooperation, but in a reciprocal exchange structure, to evaluate if reciprocal exchange led to timelier and greater exchange of resources. Results show that the reciprocal exchange structure increased flow of staff resources compared to negotiated exchange, leading to higher joint scores. Participants' cooperation also differed depending on the level of agent cooperation. In negotiated exchange, participants provided more resources to a high-cooperation agent compared to a low-cooperation agent, and in reciprocal exchange, participants provided timelier resources to the high-cooperation agent compared to a low-cooperation agent. This work departs from the typical focus on supervisory control automation and automation performance in terms of reliability, and suggests cooperative control automation and automation performance in terms of collegiality is an important area of research for human-automation interaction that enhances system resilience.

Acknowledgements

I must first thank the people who inspired me to look into this field, in order of appearance, Drs. Ying Lo, Alex Kirlik, and my HFID colleagues at Baxter circa 2010-2011, especially Dr. Ed Halpern. Thank you Ben for suggesting I look into the program at your alma mater. To my recommenders, Abigail, Joe, Agnieszka, and Brian, I hope I have shown that your effort and faith in me was not misplaced.

To my friends, especially those I made in Madison, I am indebted to your selfless effort to help me realize life balance. My labmates past and present inspired me to be a better scientist and person, as did the undergraduate students with whom I worked closely: Kim Le, Anousone Bounket, Qiwen Shu, and Tianshuo Su. Many UW-Madison faculty outside my program were important sources of inspiration, especially Drs. Montgomery, Coen, Nathan, and Gunther.

It is hard to over-emphasize how grateful I am to my committee members, Drs. Carayon, Wiegmann, Mutlu, and Werner for their time and feedback on this work, their mentorship and investment in my success. I am especially grateful to Dr. Enid Montague for many early opportunities she entrusted me, and last but not least my committee chair, Dr. John Lee. In the face of research doldrums and frustration, John was an unending source of encouragement, and he shared his ideas generously. His scholarship and example will continue to influence my research and comportment for longer than these past three years.

Finally, my decision to pursue a PhD would not have happened without financial support from my parents, who worked hard and saved in my earluy life so I could attend public university without accruing debt; the Graduate Engineering Research Scholars program; the NSF Graduate Research Fellowship Program (Grant No. DGE-1256259); and the Emerson Electric fund which made it possible for me to attend conferences and study the topic of my dreams.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

As automation capability and autonomy increase, their human counterparts become less like operators, supervisors or monitors, and more like peers in equal-authority roles. Concurrently, automation is increasingly part of dynamic work processes and is integrating into our complex human world, coordinating jointly with us on shared goals. In such circumstances, human-automation systems will more likely face coordination situations that cannot be predetermined. In joint tasks, human-automation coordination relies on cooperation because of our limitations both in knowing the right information for swift action, and in predicting the future. Whereas coordination is managing dependencies and therefore largely tactical, cooperation is predominantly social and involves the willingness to work toward the shared goal at the risk of perceived or actual individual cost. Previous research focuses on perceptions of reliability and dependability in supervisory control automation, rather than social factors influencing exchanges between more equal-authority human and machine counterparts. Because of this, little is known about how social exchange factors can influence human-automation cooperation, and subsequently adhoc coordination and adaptation to unexpected events. The goal of this research is therefore to assess how social exchange factors, such as automation behavior, work environment factors, and interaction structure, can influence cooperation in human-automation coordination and system resilience.

1.1 Advances in Automation

Advancements in automation are changing the way people interact with automation, with implications for system safety and system design. These advancements include machines that can dynamically learn from human counterparts, with the potential to work independently on tasks previously carried out by people. Rather than levels of a function, today's machines can perform whole functions, acquiring information to implementing action (Parasuraman, Sheridan, & Wickens, 2000), demonstrating a degree of autonomy that is only

recently possible. Such autonomy is demonstrated in the high-frequency trading algorithms in financial securities trading that interact with human traders (Zhang & Adam, 2012), and the pokerbot capable of alternating between aggressive and passive strategies, known as "playing the player" (Bowling, Burch, Johanson, & Tammelin, 2015; Kaplan, 2013). Other advances include algorithms that analyze social situations to coordinate with people in more dynamic situations (Wagner & Arkin, 2008), and automation architectures that strive to achieve more natural interleaving between people and automated systems (Allen, Guinn, & Horvitz, 1999). Machines that can engage in sophisticated game-theoretic reasoning, cooperating or competing, and actively coordinating with human counterparts, demonstrate abilities that go beyond traditional rule-based automation.

As automation becomes more like equal-authority teammates, and the heterogeneity or autonomy of agents in an organization increases, group structures flatten. In these more lateral structures that lack pre-defined coordination mechanisms, it is more common to have multiple simultaneously active goals than in supervisory-control situations. Practitioners therefore must be able to reconcile potentially conflicting goals (Castelfranchi, 1998; Woods & Hollnagel, 2006). For function allocation, this type of coordination relationship has been described as mixed-initiative (Horvitz, 1999) or interactive control (Van Wezel, Cegarra, & Hoc, 2011). While mixed-initiative refers specifically to characteristics needed to achieve interleaving of joint action, interactive control refers generally to a relationship where the human and automation are both involved in an activity and each may propose, check, or evaluate solutions or partial solutions. Both differ from supervisory control, a "hierarchical control scheme whereby a system…having sensors, actuators and a computer, and capable of autonomous decision-making and control over short periods and in restricted conditions, is remotely monitored and intermittently operated directly or reprogrammed by a person" (Sheridan & Verplank, 1978, p. 1-1).

As defined, supervisory control automation reflects a different approach to automation integration in human work systems, and has led to much research on the costs of such an approach (Kirlik, 1993), including issues of reliance and compliance (Meyer & Lee, 2013) and maintaining situation awareness (Endsley & Kaber, 1999; Endsley, 1996). While research on understanding how to prevent failures in existing systems is still critical, increasingly, it is recognized that supervisory control automation may not be the best solution for dynamic coordination in high-criticality work domains. When the opportunity for human intervention at the point of automation failure is limited in time, and the costs of system failure are high, the principles behind lateral coordination at least acknowledge that coordination emerges in unforeseen circumstances. Furthermore, by virtue of its relationship structure, lateral coordination engages the human counterpart, sidestepping to some degree issues associated with a monitoring role, including automation complacency and skill degradation (Bailey & Scerbo, 2007; Miller, Funk, Goldman, Meisner, & Wu, 2005; Parasuraman, Molloy, & Singh, 1993).

To coordinate laterally with people in the natural world, automation needs to be able to reconcile goals, communicate naturally, and sustain an interactive dynamic (Allen et al., 1999). An interactive dynamic, where either person or automation may take initiative, allows each agent involved to negotiate its role opportunistically, to best address the problem at hand, depending on the circumstances and the skills each contributes best. Previous research in human-automation interaction focuses primarily on information-processing approaches and human operators' reliance on or compliance with automation for varying levels of automation performance. In these new automation contexts, with increasing autonomy and applications in more dynamic environments, social exchange factors that go beyond reliance and compliance may play an increasingly prominent role. These factors include signals from an automated agent's behavior, work environment demands that influence those behaviors,

and the social exchange structure of human-automation interaction. While a gross simplification of the potential variables involved, these three dimensions focus on the key mechanisms impacting the analytical-based and affective-based judgments in cooperation that enable effective lateral coordination. In the face of incomplete information, reconciliation of laterally competing goals – even with automation – will rely on trust and social processes (Lee & See, 2004; Parasuraman & Miller, 2004; Wagner & Arkin, 2011), and thus the automation's contribution to system performance may be more defined by collegiality than by reliability. Automation collegiality, in turn, cannot be assessed in a vacuum when real world coordination tasks involve more than establishing positive relationships (Muir, 1987; Parasuraman & Miller, 2004).

1.2 Cooperation and Resilience Engineering

In leveraging social-exchange theory to better understand human-automation cooperation in lateral relationships rather than supervisory relationships, a central contribution of this work is to Resilience Engineering. Resilience Engineering is an approach to system evaluation and design that considers how jointly interacting agents can work to achieve global goals in increasingly dynamic and safety-sensitive environments (Hollnagel, Woods, & Leveson, 2006). Though Resilience Engineering has yet to provide useful quantification methods to help mitigate risk and accelerate recovery from unexpected events, it rightly emphasizes a paradigm to help avoid the brittleness of traditional engineering approaches that focus on reducing variability and predicting based on past events (Sheridan, 2008). Rather than measuring productivity and performance simply by the quantity or quality of widgets processed, Resilience Engineering focuses on processes that help organizations anticipate, mitigate, and prepare for graceful recovery from adverse events, include maintaining a safety margin by sharing resources (Stephens, Woods, Branlat, & Wears, 2011), and focusing on system-level consideration of outcomes rather than local or

individual-level consideration. Collaborating to achieve global goals may require compromise of individual local goals, which can be influenced by social exchange factors. For automated systems interacting with human counterparts, effective collaboration requires an understanding of how people cooperate with automation, and the factors influencing those decisions.

1.3 Research Questions and Scope

The purpose of this research is to consider how elements of social exchange that define human-automation interaction enable effective lateral coordination. This approach is motivated by the trend towards increasingly autonomous automation partnering with humans as peers in dynamic joint tasks, and a lack of consideration of these factors in current human-automation interaction research. To achieve this objective, this work explores the following overarching hypotheses:

1. Automated agent behaviors that signal varying levels of cooperation will affect participants' propensity to cooperate when coordinating on dynamic a joint-task.
2. Changing the social exchange structure of human-agent interaction, from a structure that emphasizes consideration of others' needs, rather than self needs, will result in enhanced cooperation through appropriate resource giving.

Considering lateral coordination from the perspective of social exchange factors raises many issues and connotes a substantial research agenda. Not included in this research, but important for its application, would be to consider what type of control relationship is more appropriate for a given task and domain, e.g. manual, supervisory, advisory, interactive, or fully automatic (Kirlik, 1993; Van Wezel et al., 2011); to what extent tasks need to be interdependent (e.g. Cummings, 1978); and how to determine goal prioritization (examples in Kelley et al., 2003). This dissertation derives inspiration from a wide range of literature for an enhanced understanding of human behavior and responses to an autonomous agent.

1.4 Practical Contributions

Drawing from the trend of increasingly autonomous automation and insights on interaction from social exchange theory, the current research aims to address three aspects of human-automation interaction: interaction strategies in a joint task, changes in the work environment that require goal adjustment, and the structure of interaction that affords different social mechanisms of cooperation. Having a better understanding of how people respond to different automation behaviors can inform design of automation that elicit cooperative engagement and appropriate action in a joint task. Furthermore, understanding how changes in the work environment interact with agent behavior and people's response to both the agent and their environment can clarify how tradeoffs in the environment and in resources impact cooperation. Finally, testing different social exchange structures will turn attention to alternative design strategies in the space between cooperating entities, and highlight the potential role of interaction structures on cooperation. Focusing on these aspects contributes to the consideration of social exchange components of human-automation interaction design, and mechanisms of cooperation for human-automation teams and organizational resilience. This work also contributes to the empirical basis of Resilience Engineering by demonstrating how social exchange factors are important for the design of organizational processes and technology that lead to graceful extensibility through the benefits of joint work (Woods, 2015)

1.5 Theoretical Contributions

This is the first known study to use a microworld environment to explore interdependent social exchange factors of human-automation cooperation in a dynamic joint-task environment. This research departs from research on cooperation between people mediated by reliance on automation (Gao, Lee, & Zhang, 2006), as well as the typical information-processing approach to human-automation interaction (Parasuraman et al., 2000). Instead, it

takes a relational, social exchange perspective that focuses on automation collegiality and joint cooperation, rather than on automation reliability and operator reliance. The assessment of human-automation interaction in this context is critical to future automation design and research, as automation increases in dynamic and safety-critical environments. Findings from this work may also extend to other interactions between humans and technology, or between remote team members. Mainly, this work serves as a starting point for an enhanced understanding of the fundamentals guiding behavior in human-automation systems that more lateral relationships make relevant.

Chapter 2 Background and Literature Review

This chapter provides the motivational backdrop of this research. It presents background information and focuses relevant concepts informing the hypotheses explored. First, examples of automation advances are provided as justification for a closer look at human-automation interaction, specifically the flattening of their work relationship. These trends in technological advancements and organizational restructuring are then connected to the broader view of Resilience Engineering and its relevance to system design in high criticality environments. Previous work on human-automation interaction and related constructs such as trust and coordination are then reviewed in the context of Resilience Engineering, in particular human-agent cooperation and achieving joint tasks. Following this discussion, the social exchange worldview used is presented to clarify and focus key concepts driving the hypotheses explored. Finally, the significance of this research is summarized followed by a statement of the research objective and questions addressed.

2.1 Advances in Automation

From self-driving vehicles to expert decision-support systems and robots that learn, advances in automation have led to machines capable of automating whole functions that previously required human intervention. Furthermore, adaptable automation means people are increasingly able to influence and test automation. In such cases, the core of automation performance may actually rely on human input, changing the nature of the human-automation relationship from supervisory control to advisory control or interactive control (Van Wezel et al., 2011). Rather than determining the types or levels of automation for separate system functions (Parasuraman et al., 2000), trends in automation and artificial intelligence are showing approaches to automation that support a more interactive relationship with human counterparts (Allen et al., 1999; Breazeal & Scassellati, 1999; Castelfranchi, 1998; Knight, 2013; Wagner & Arkin, 2008). These approaches reflect both technical advancements, as

well as a vision to apply automation in increasingly dynamic, and increasingly complex environments.

2.1.1 From supervisory control to interactive partner

In supervisory control automation, the human operator is often relegated to monitoring automation performance, and is tasked with intervening if the automation fails. System performance under this arrangement thus depends on the interplay between the automation's reliability, the human operator's ability to assess its reliability and to appropriately rely on or comply with the automation (Meyer, 2004; Parasuraman & Riley, 1997). However, a combination of increasingly capable automation and increasingly complex systems suggests a need to go beyond the supervisory control paradigm. Paradoxically, the work demands intended for increasingly capable automation require collaborative relationships with people, where successful interactions are synonymous with cooperation and teamwork rather than with consistency and reliability. Under this arrangement, the potential for goal conflict and ambiguity increases with more interactive automation, outlining a different human-automation relationship in which social factors play an increasingly central role (Lewicki, McAllister, & Bies, 1998).

Examples of such automation include the recent advances in commercial automotive engineering, social robots, computational trading algorithms, and uninhabited vehicles. These developments place automation in increasingly interactive situations, communicating and coordinating with people in real-world circumstances. Along with these technological trends is a need for automation in more social and functional roles, such as gym trainer or caretaker (Goetz, Kiesler, & Powers, 2003). Machines have already demonstrated a capability to take on roles in many knowledge-based tasks and social tasks, (Ferrucci, 2012; M. Kaplan, 2013; Levy, 2012) that were partially or fully fulfilled by people. However, such machines are largely in domains where unreliability does not pose as severe of costs as it could in other

domains where timing is more critical. As automation expands its scope into dynamic roles within safety-critical domains, where coordination emerges from multiple cooperating entities rather than derives from pre-planned processes (Casper & Murphy, 2003; Defense Science Board Task Force, 2012; Robinette, Wagner, & Howard, 2013), current approaches to human-automation integration will be challenged.

2.1.2 Agency and attribution

Beyond technical capability, another reason for reconsidering social factors influencing human-automation interaction is the development of increasingly embodied and socially aware agents (Adams, Breazeal, Brooks, & Scassellati, 2000; Lee, Knox, & Breazeal, 2011; Verberne, Ham, & Midden, 2012). Such features and capabilities that display agent-like behavior may induce unintended social responses (Nass & Moon, 2000). Particularly with automation actions and decisions that can signal intention to others, people's tendency to ascribe motives to objects may affect their decisions to cooperate with them (Kelley, 1973).

There is also a trend of increased automation embodiment and more socially-aware automation, such as robots and avatars, bolstered by research on voice (Nass & Brave, 2005), touch (Markussen, 2009), conversation (Cassell & Bickmore, 2000), and goal-oriented movement (Hoffman & Ju, 2014; Mutlu, Kanda, Forlizzi, Hodgins, & Ishiguro, 2012; Terada, Shamoto, Mei, & Ito, 2007). These machine abilities with embodied applications reflect a desire to bring automation into our more dynamic human world (Brooks, 1991; Horvitz, 1999; Wagner & Arkin, 2008), with designers leveraging people's tendency to interact with things as if they had agency (Epley, Waytz, & Cacioppo, 2007; Takayama, 2009). Automation working in social environments moves them from exhibiting relatively understandable behaviors, i.e. teleoperation where behaviors map directly to function, to the more sophisticated and ambiguous behaviors of social interaction, a complexity enabled by

agency and embodiment rather than (more simply) capability or performance in controlled, well-defined settings. Cooperation with such automation may thus depend on both inadvertent and purposeful social cues of new machines.

The attribution of humanlike properties, characteristics, or mental states to real or imagined nonhuman agents and objects goes beyond descriptions of imagined or observable actions (e.g. "the robot is sad"). At its core, it is inference about unobservable characteristics and motivations of a nonhuman agent that increases one's ability to make sense of an agent's actions, reduces uncertainty associated with an agent, and increases confidence in predictions of this agent in the future (Epley et al., 2007). This reduction of uncertainty also fulfills the human desire to interact effectively with their environment (White, 1959). While predicting the extent to which certain anthropomorphic characteristics elicit social responses is beyond the scope of this research, there is evidence supporting the general idea that experiencing increased agency in non-human objects or entities could lead to increased attribution of human-like qualities (Benninghoff, Kulms, Hoffmann, & Krämer, 2012; Fussell, Kiesler, Setlock, & Yew, 2008; Nass, Steuer, Tauber, & Reeder, 1993; Takayama, 2009).

The concept of agency has been ill-defined, but has been used to refer to the ability of an entity to act in an environment on its own (Castelfranchi, 1998; Franklin & Graesser, 1997). However, this definition is similar to autonomy, which has been defined as "the amount of freedom and discretion an individual has in carrying out assigned tasks" (Langfred, 2007, p. 886), or referred to as acting "in pursuit of [one's] own agenda" (Franklin & Graesser, 1997, p. 26). These definitions assume not only the ability of action, but also imply an assignment of ability, both of which are properties of the entity. This is consistent with Bradshaw et al.'s (2004) discussion of self-sufficiency (descriptive of the entity) and self-directedness (prescriptive onto the entity) as two dimensions of autonomy. In their spectrum of autonomy, autonomy is closely related to the ability to reconcile multiple

competing goals – including the local goals relating to self-sufficiency on a particular task, and the shared goals relating to the ability to interact flexibly with other entities. This definition is necessarily context dependent. For example, a light-sensing robot will have little autonomy in the context of a dark room, because its environment hinders its abilities. Similarly, a self-driving vehicle might be considered to have a higher degree of autonomy if it were operating in a driving environment that did not involve other people.

A more recent description of agency considers both the properties of the entity, as well as the perception of the entity. In this definition, agency refers to entities that are perceived and responded to in-the-moment as if they were agentic despite the likely reflective perception they are not agentic (Takayama, 2009). This definition differentiates reflective beliefs (e.g. "I know this robot is just an object") from in-the-moment responses that treat the object as if it had a high degree of agency (Takayama, 2009), and supports the idea that reflective perceptions are fundamentally different from perceptions during interaction, in a biological sense (Schilbach et al., 2013). Thus, while autonomy refers mostly to the objective ability of an entity to accomplish functional goals on its own, agency includes both the ability and subjective perception of ability during interaction. This is an important distinction for automation that will interact with people, because it is perceptions and subjective realities that people judge and act upon (Tversky & Kahneman, 1974).

Despite the excitement over advances in machine autonomy and future robotic teammates, Groom and Nass (2008) posit that imbuing robots with the "humanness" assumed of human teammates will be extremely challenging because people's innate expectations for team-appropriate behavior cannot be fixed with technological innovation. Establishing trust between people and agentic automation may be one of the most daunting problems for the development and success of such mixed teams, particularly to achieve the type of cooperation needed to coordinate in environments with multiple competing goals. To work toward this

ideal, Groom and Nass (2008) observe that future research intent on developing human-robot teams must go beyond technical performance of automation to address the social and organizational qualities of successful teaming.

2.2 Resilience in Sociotechnical Systems

The interest in mixed teams may be partially driven by the demand for more flexible systems, in place of more traditional systems that strive to reduce variability. Traditional engineering approaches such as Reliability Engineering, which compares system performance to predetermined criteria (Billinton & Allan, 1983), and Lean Engineering, which tries to improve output by reducing variability (Main, Taubitz, & Wood, 2008), can lead to brittle systems and catastrophic failure in complex, safety-critical settings (Nemeth, Wears, Woods, Hollnagel, & Cook, 2008). Therefore, most complex, safety-critical settings still need people to fill the gaps that automation creates or that system designs cannot predict and anticipate. In complex, safety-critical settings the idea of more flexible coordination is increasingly popular, as automated systems improve and become a part of such settings, actively coordinating with human counterparts. Resilience Engineering is one such perspective that promotes the need for more flexible systems (Hollnagel, Woods, & Leveson, 2006).

Woods (2015) clarifies the different existing notions of resilience in the safety engineering literature: Resilience is graceful extensibility and refers to the ability to manage sustained adaptability of a layered network system; this differs from resilience as rebound, to restore a system to previous conditions prior to a disruption; or resilience as robustness, having an expanded set of models that effectively respond to disturbances. Rather, the concept of resilience as graceful extensibility derives from combining "graceful degradation" – the ability of a machine or networked system to maintain limited functionality, even when a large part is inoperative, to prevent catastrophic failure – with

software “extensibility” – the ability to have new functionality extended without affecting, or minimally affecting, existing system functions (Woods, 2015). An application of graceful extensibility to automation design considers how, as part of an interacting network of agents, automation can cooperate and draw on shared resources to accommodate surprises.

2.2.1 Organizational restructuring

The movement of automation towards more specialized functions and lateral coordination with human counterparts is not just a reflection of technological advances, but also a way to maintain advantage in a competitive market. Following the Industrial Revolution, increased job specialization for human workers led to finer-grained division of labor and the subsequent need to coordinate among specialties (Grant, 1996). This disaggregation of work from Tayloristic hierarchical structures of the Industrial Revolution to more interdependent lateral work processes of the Information Age (Adler, 1997, 2001; Blomqvist & Stahle, 2004), led to achievements previously thought impossible: reviving patients from death-like states, erecting complicated skyscrapers, and landing military aircraft on carriers at sea (Gawande, 2009; Rochlin, La Porte, & Roberts, 1987).

While the aforementioned achievements of lateral work processes refer to predominantly human organizations, automation is likely to continue taking on jobs that were once the purview of human workers (Frey & Osborne, 2013). Alongside these organizational changes, new approaches to human-automation interaction, such as interactive function allocation (Pritchett, Kim, & Feigh, 2013; Van Wezel et al., 2011), mixed-initiative interaction (Allen et al., 1999), and social situation analysis in robots (Wagner & Arkin, 2008) suggests that more lateral approaches to human-automation system design are within reach and promising. Implied in these approaches is a level of autonomy individual agents have within the larger networked system, similar to the autonomy of human workers in self-regulating work groups (Cummings, 1978).

Lateral work structures tend to be more useful when external controls, including hierarchical work structures, are unable to reduce uncertainty. Thus, increasing the internal control of individual workers becomes key for successful coordination. Internal control is having authority to work autonomously, but often collaboratively, addressing issues as they arise instead of following top-down procedures far removed from the issue. In these types of work systems, coordination and cooperation emerge more than they are predetermined. As advanced automation become capable of coordinating dynamically with human counterparts, performing reliably and robustly on predetermined criteria may be less critical than flexibly adjusting alongside human counterparts as conditions evolve (Woods, 2015; Zieba, Polet, Vanderhaegen, & Debernard, 2010).

2.2.2 Flexible systems in dynamic environments

For supervisory control automation, joint performance depends on automation reliability, the human operator's ability to assess automation reliability, and appropriately rely on or comply with the automation (Parasuraman & Riley, 1997). However, for more advanced automation in volatile environments, interactive control, sometimes referred to as collaborative control, allows both agents to combine their competencies and negotiate conflicts (Zieba, Polet, Vanderhaegen, & Debernard, 2009). Rather than efficient but inflexible performance, interactive control supports resilience through adaptation to novel situations, accommodation of bugs in the system itself, and recovery from errors in following procedures (Woods, Johannesen, Cook, & Sarter, 1994). Interactive control, now more possible and more likely due to the aforementioned technological advances, is similar to teamwork in that it allows a group of agents to adapt collectively to unexpected perturbations in the work environment (Zieba et al., 2009, 2010). Such perturbations have the potential to affect work systems outcomes, manifested in the interactions between agents.

While Woods (2015) discusses a larger system of networked agents more generally, Van Wezel et al (2011) focuses on the human-automation dyad. Van Wezel et al's (2011) characterization of dynamic versus static scheduling tasks for human-automation function allocation supports the idea that demands of the system should inform the type of human-automation relationship, which range from advisory to interactive and supervisory. This characterization of the task environment is important for determining the types of human-automation relationships that would best address such environments. From management science to resilience engineering to mixed initiative approaches, there seems to be cross-discipline support that in more unpredictable environments, function allocation needs to move beyond predefined roles and responsibilities (Allen et al., 1999; Baker, Day, & Salas, 2006; Castelfranchi, 1998; Cummings, 1978; Zieba et al., 2009). Instead, successful coordination in these environments relies more on the ability to collaboratively resolve multiple goals.

When cascading events push operations in a way that stretches an organization's response to increasing demands, individuals and groups make successful adaptation possible by adjusting strategies and resources to provide the additional capacity (Nemeth et al., 2008). Such adjustments in and of themselves may increase workload, particularly in the ad hoc judgment and decision-making required, to make the appropriate choices given the environment, to best respond to requests from others, and to proactively consider the shifting needs of others and the organization as a whole. The potential of automation to introduce new forms of workload in coordinating with people may lead to breakdowns in coordination and cooperation within the organization (Kirlik, 1993; Mutlu & Forlizzi, 2008; Woods, Tittle, Feil, & Roesler, 2004).

These new work environment contexts for automation highlight the limitations of the Levels of Automation framework (Parasuraman et al., 2000), mainly that in dynamic tasks

the time between acquiring information and implementing action cycles more quickly (Van Wezel et al., 2011), and it is often difficult to predict the inputs in certain work environments (Cummings, 1978). Designing for resilience suggests a new approach is needed, for understanding potential interactive control issues, as an important complement to current supervisory control frameworks.

2.3 Human-Automation Coordination

2.3.1 Are they teammates?

Teams have been defined as a set of two or more individuals interacting adaptively, interdependently, and dynamically toward a common and valued goal, and they differ from groups in terms of task interdependency, structure, and time span (Salas, Burke, & Janis, 2000). As the line blurs between human-automation teams and automation-mediated human-human teams, research on teamwork may be a source for better understanding how interacting autonomous agents coordinate effectively. Such work has considered how teamwork is an essential component of avoiding adverse events in complex, hazardous environments where error consequences are high but their occurrence is low (Baker et al., 2006), how human teams can learn to adapt to novel situations through perturbation training (Gorman, Cooke, & Amazeen, 2010), and how implicit and explicit communication, coordination strategies, and cooperation involving trust and team cohesion are key components of teamwork (Salas, Wilson, Murphy, King, & Salisbury, 2008).

While human-automation interaction researchers have already considered studying human teams to better understand how people coordinate and cooperate with others (Allen et al., 1999), there are limitations to taking direct parallels from the teamwork literature comprising human members and applying them to mixed teams. A large component of teamwork literature refers to Shared Cognition, or the idea that teammates can share mental models, situation awareness, and a common language for communication. Shared Cognition

is the theoretical basis for understanding how teams adapt performance processes and interpret environmental cues to make compatible decisions and to coordinate (Salas, Cooke, & Rosen, 2008). Groom and Nass (2007) argue persuasively that because robotic teammates lack humanlike mental models and a sense of self, and that mental models are so fundamental to human cognition, such machines may be rejected as members of human teams. It is still unclear the extent to which this human quality of sharing mental models are assumed of teammates.

More recent methods applied to the study of teams questions whether measurements of shared knowledge or shared understanding is predictive of team performance (Cooke, Gorman, Myers, & Duran, 2013). Therefore, an "interactionist" approach has been suggested to better understand team performance through communication and coordination process measures rather than static performance measures (Cooke & Gorman, 2009). Such approaches that focus on process measures may be more relevant for human-automation interaction as well, since it avoids the need for the automation to have a mental model at all, and instead focuses on the dynamic mechanisms rather than static factors that lead to coordination or cooperation.

Another limitation of the teamwork literature is the conceptualizations of coordination and cooperation that largely derive from descriptive observation of human teams rather than from a first principles perspective. A definition of cooperation as "the desire to coordinate by engaging, anticipating, and predicting one another's needs" (Salas, Wilson, et al., 2008) is difficult to separate from concepts like motivation and coordination, and more or less ignores the idea that certain behaviors associated with cooperation may not always lead to the desired system-level outcomes. Work on human teams can be an important source of inspiration, to better understand how people respond to others' actions, and the desired processes and outcomes of good team coordination. But it is difficult, and may be unnecessary to build

automation that appropriately reflects human mental models in all situations. It may be sufficient to develop automation that can recognize situations that demand trust and cooperation without relying explicitly on a shared understanding with humans (Wagner & Arkin, 2011), and to better understand the interrelationship between processes in teamwork (Salas, Cooke, et al., 2008). In the meantime a more specific definition of cooperation is needed to clarify what is meant by human-automation teaming, and to clarify how cooperation is related to system resilience.

2.3.2 Coordination and cooperation

Drawing from the integrated work that forms coordination theory, coordination is defined as dependency management (Malone & Crowston, 1994). The intuition is that without interdependencies, there would be nothing to coordinate. One example of a common dependency is coordination of resources among workers. The process for managing such a dependency may be using the "first come, first serve" priority order. Other processes that help manage coordination include notifications, standardizations in usability design, and task decomposition for work design. When people are involved, however, understanding coordination also involves understanding their incentives, motivations, and emotions due to their impact on decisions and actions, particularly when their goals conflict with others. To the extent that coordination emerges in lateral team structures without being planned, cooperation plays a greater role in achieving a joint outcome. Cooperation is most easily identified in actions, that despite individual costs, benefit a group of two or more members (adapted from Dugatkin, Mesterton-Gibbons, & Houston, 1992).

The importance of cooperation in resilience is highlighted in the idea of a margin of maneuver—a cache of actions and resources that allows the system to function despite unexpected demands (Woods & Branlat, 2010). Systems without an adequate margin become brittle and are unable to withstand unexpected demands. Based on studies of emergency

department interactions with other units in a hospital system, Stephens, Woods, Branlat, and Wears (2011) identify three classes of creating or maintaining a margin of maneuver: defensive strategies, autonomous strategies, and cooperative strategies. Defensive strategies increase the margin of an organizational unit by restricting another unit. An example of this would be actions associated with greediness, such as borrowing resources from another unit in anticipation of oncoming demand, without consideration of the other unit's needs. Autonomous strategies involve local focus and may involve reducing interactions with other units because the benefits of sharing resources are not believed to be relevant or worth the costs. Such costs include both the value of the resources shared, as well as the individual time and effort involved in coordinating with another entity, including cognitive demands like gathering and processing information about the other's needs juxtaposed to self needs and global needs, and choosing the best action. Cooperative strategies involve the effort of two or more units that, through coordinated and collective action, recognize or create common-pool resources from which both units can draw. Such actions help avoid the tragedy of the commons when resources are limited (Ostrom, 1999), and may enhance system resilience by giving the overall network a larger pool of resources to draw upon for greater margin of maneuver.

Agents' priorities may thus influence people's decisions to cooperate, such as when an agent's actions or other signals are perceived as the agent prioritizing individual interest over the shared group interest, so the costs of coordinating may not be justified. Just as initial trust is important for long-term collaboration between people, automation, and organizations (Li, Hess, & Valacich, 2008; Zheng, Veinott, Bos, Olson, & Olson, 2002), initial cooperative action has been found to be a more successful long-term strategy than an initial uncooperative action (Axelrod & Hamilton, 1981). Hoc (2001) identifies criteria that define situations where cooperation is relevant: each agent strives toward goals; can interfere with

the other's goals, resources, procedures, etc.; and tries to manage the interference to facilitate individual activities and the common task when it exists. Similar to the teamwork literature, Hoc defines cooperation without explicitly including goals. However, this work considers goals to be inseparable from cooperation. Goals are a central component for explaining how cooperation facilitates lateral coordination. Without well-defined roles or responsibilities between people and automated agents, the potential for people to act in ways that tend to advance personal goals rather than the joint goals, or their partner's goals more locally, are a fundamental difference from problems where coordination is the focus.

The definition of cooperation used in this dissertation is the act of compromising individuals' goals and self-interest to achieve shared goals and collective interest. This definition arises from what Clark (1996) calls private and public goals in social psychology, Ostrom (2000) calls self and collective interest in economic behavior, Hoc (2001) calls ego-centric and collaborative interests in multi-agent interaction, and Woods (2004) calls local and global goals in resilience engineering. While these constructs are historically and thus conceptually distinct, they share the idea that human cooperation emerges through reconciliation of myriad competing goals through social processes. In such situations where interdependent activity involves shared resources and conflicting goals, cooperation is essential for coordination. Thus, rather than collaborative control, which often refers to managing functional dependencies by adapting levels of autonomy, cooperative control may be an appropriate term for scenarios where cooperation is needed for achieving shared goals that conflict with individuals' goals.

Figure 1 illustrates the difference between a pure coordination scenario and a cooperation-centric scenario, using a dyad and formalisms from game theory for simplicity.

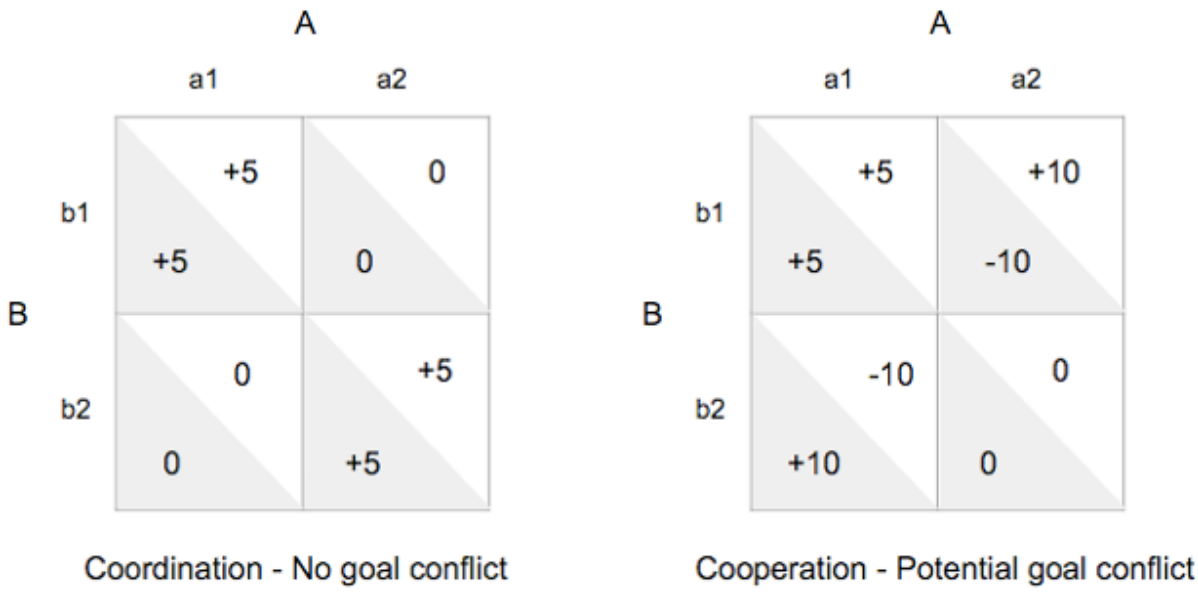

Figure 1. The left matrix shows two agents' interest in the same combinations of behavior, i.e. coordination. The right matrix shows potential goal conflict, and how trust is needed, i.e. cooperation.

The matrix on the left, adapted from the "Corresponding Mutual Joint Control" (Kelley et al., 2003, p. 160), shows two agents' common interest in the same combinations of behavior; the best outcomes simply depend on agent A and B synchronizing or arranging their behavior so that both choose the same option, option 1 or option 2. The cooperation matrix, adapted from "Matrix Representation of the Single-Trial Prisoner's Dilemma" (Kelley et al., 2003, p. 189), shows how potential goal conflict means a component of trust is needed (that the other will not choose to defect) for both parties to avoid negative outcomes. Cooperation is thus about arranging actions in the desired configuration, and if individual and joint goals differ then aligning them.

It should be acknowledged that human-automation roles are not often as concrete as this representation might suggest. Decisions often comprise multiple sub-decisions with myriad interacting variables. There are certainly limitations to the extent that real world situations, roles, and decisions can be represented so simplistically. But because perceived alignment of goals can play a critical role for interacting human and automation agents, cooperation seems better conceptualized as a sub-class of coordination that incorporates goal

dependencies. Goals are defined as a mental representation of a desired endpoint or target situation that can: guide action by repeated tests of the action against the representation itself, control the action search and selection, and qualify its success or failure (adapted from Castelfranchi, 1998). Since a goal-directed agent may have more than one goal active in the same situation, and skills and resources are limited, the agent must be able to form choices regarding prioritization of goals, unable to achieve all of them. That the goals of multiple agents are dependent on one another is the nature of a joint task.

Goals, not behavioral differences, are said to distinguish social action from non-social action, forming the basis of exchange and cooperation among interacting, intelligent agents (Castelfranchi, 1998). The increasing agency demonstrated in new automation, coupled with the more informal interactions of lateral coordination may mean the success of future systems comprising people and autonomous agents may depend increasingly on social constructs like trust. The intuition that trust is needed for cooperation in joint decision-making is supported by studies on human-automation interaction which show having shared goals is associated with higher levels of trust in automation (Cramer et al., 2008; de Visser & Parasuraman, 2011; Verberne et al., 2012). In light of increasingly autonomous automation and the need to engineer resilient systems, better understanding the factors influencing human-automation cooperation seems crucial. System resilience demands alternatives to the more explicit procedures of the supervisor-subordinate relationship, and these alternatives include informal interactions that promote coordination and cooperation (Rochlin et al., 1987).

2.3.3 Trust and cooperation

Trust, a substitute for formal controls, guides human behavior toward automation when a complete understanding of the automation is difficult or impractical (Lee & See, 2004). Though cooperation may exist in the absence of trust, such as when controls are in place and the situation does not put a party at risk (Mayer, Davis, & Schoorman, 1995), trust

can reduce transactional costs in obtaining cooperation when goals conflict (Kramer, 1999). Especially when problems cascade, demands for cognitive activity increase as do the demands for coordination across agents (Woods & Patterson, 2001). In extreme cases, people may be even less likely to interact with automation while experiencing high stress with physical and cognitive fatigue as was the case for search and rescue teams at the World Trade Center (Casper & Murphy, 2003). These increased demands may accentuate bonding or fracturing of teams. Without trusting that cooperation will lead to better long term outcomes despite short term costs, people may resort to defensive strategies to preserve local gains at the cost of the global system (De Dreu et al., 2010; Stephens et al., 2011). In the face of transaction costs, trust is ubiquitous (Arrow, 1974).

Trust in supervisory control automation has been said to be different from interpersonal trust because it is unilateral rather than relational (Lee & See, 2004). However, recent work focusing on more autonomous automation (e.g., agents or robots) suggests a relational view of trust may apply (Bray, Anumandla, & Thibeault, 2012; Fink & Weyer, 2014; Wagner & Arkin, 2011). A relational perspective of trust may provide a better basis for deriving principles of interaction design and cooperative control than an information processing perspective. For example, a relational perspective that considers social exchange theory can help explain why people cooperate even when it is not in their best interest to do so. Social exchange theory accounts for the idea that communication with a partner shapes perceptions of an exchange partner and interpretation of the partner's subsequent actions (Bottom, Holloway, Miller, Mislin, & Whitford, 2006). It also accounts for the idea that people are generally good at detecting the trustworthiness of an exchange partner based on such signals (Janssen, 2008). These perspectives that view trust as a relational construct, when the unit of analysis is at least two, rather than as an information-processing construct

where the unit of analysis is one, may highlight important components of cooperation with automation that the trust in automation literature largely ignores.

2.3.4 Trust in automation

Previous work on trust in supervisory control automation, which refers to a more hierarchical relationship between people and automation, tends to focus on reliance and perceptions of performance. Most notably, Lee and See's (2004) review is cited in domains as diverse as the military (Barnes et al., 2014; Bertuccelli & Cummings, 2011; Lyons & Stokes, 2011), healthcare (Ho, Wheatley, & Scialfa, 2005; Montague, Kleiner, & Winchester III., 2009), and driving (Kazi, Stanton, Walker, & Young, 2007; Verberne, Ham, & Midden, 2015) among others, for its argument that trust guides reliance on automation. Reviews following Lee and See (2004) include Madhavan and Wiegmann (2007), which focuses on trust in decision aids highlighting differences between human-human and human-automation trust, and a meta-analysis of trust in robots (Hancock et al., 2011) posits that trust in robots may differ from trust in other forms of automation, though trust is still largely discussed in terms of people's use or disuse of automation. These approaches neglect the relational aspects of trust that may feature more prominently with increasing automation autonomy and agency.

A more recent article by Hoff and Bashir (2014) reviews human-automation studies from 2002 to 2013, summarized as a three-layer model of human-automation trust: the human operator, the environment, and the automated system. These broad categories are described as dispositional trust, situational trust, and learned trust, and encompass a wide range of factors influencing trust and trust formation. While the descriptions of these categories hint at the interdependence of the constructs, and thus potential relational factors, the integrated findings still refer to trust in terms of reliance and use of automation. Hoff and Bashir (2014) do refer to cooperative contexts in their initial conceptualization of trust, but

the studies reviewed primarily relate trust to reliance and use of automation rather than to cooperative exchange. Key elements missing from their review – a reflection of the existing empirical research rather than the authors' approach – include reciprocity and the interdependence of factors in learned trust, ascriptions of intention and the interdependence of factors in situational trust, and the inherent interdependence within the social structures of exchange in situational trust.

2.3.5 Trusting automation agents

To realize the potential of human-automation teams that support system resilience, automation designers need a better way to parse the mechanisms that promote flexibility and cooperation from mechanisms that can lead to brittle systems and breakdowns. Automation that considers environment changes, and adjusts goals and intent accordingly, may become more prevalent in the future, and what their actions signal to their human counterparts may lead to unexpected outcomes given the lack of research in such scenarios with automation. Understanding how intention is communicated and received between people and more autonomous automation, and how such communications can lead to system-level outcomes requires consideration of human-automation dynamic interactions. At the system level, individuals often need to fill in the gaps that system design or system failures produce, and cooperation is critical in enabling people to fill those gaps. Thus, rather than approach trust in automation as an asymmetric relationship, trust in automation as a more symmetric relationship – where people provide inputs to automation that can adapt, and automation can signal to not only elicit trust, but also to repair trust – would improve our understanding of how cooperation with increasingly autonomous automation evolves over time.

2.4 Social Exchange Worldview

Social Exchange Theory has the potential to inform a new understanding of trust in automation as a more symmetric relationship. Social exchange can be thought of as a

transference of something from one entity to another in return for something else, and may involve proaction and reaction within an exchange unit, which can include two people or a person and her environment (Roloff, 1981). Whereas economic exchanges tend to be impersonal, social exchanges create feelings of personal obligation, gratitude, and trust, although many real world exchanges include aspects of both. Because social exchange relationships lack formal sanctions, or unspecified obligations (Cropanzano & Mitchell, 2005), people tend to attribute positive or negative feelings to the exchange entity or partner. Behavioral economics and other work that takes a social exchange worldview have done much to challenge our notion of people as either "rational" or "irrational" actors (Ariely, 2008; Kahneman & Tversky, 1979; Schelling, 1960; Simon, 1955). Decisions made "irrationally" may in fact turn out to be rational. In a repeated exchange situation with a long horizon, making the "irrational" choice of trusting an unknown entity, and following a tit-for-tat strategy in future exchanges, mathematically leads to the most advantageous outcome for both individuals and the larger group (Axelrod & Hamilton, 1981).

Social exchange factors between interacting entities have been studied in contexts such as interpersonal and organizational relationships. Social exchange factors that influence a partner's trustworthiness and decisions to trust the partner, include signals of intention, reciprocity of actions, level of uncertainty, and history (Ferrin, Bligh, & Kohles, 2007; Kramer, 1999; Mayer et al., 1995; Wagner, 2013). The structure of the exchange (Kelley et al., 2003; Thibaut & Kelley, 1959) and characteristics of the environment, such as its complexity, novelty, levels of risk and uncertainty have also been found to influence trust and cooperation (Delton, Krasnow, Cosmides, & Tooby, 2011; Molm, Takahashi, & Peterson, 2000; Riedl, Mohr, Kenning, Davis, & Heerkeren, 2011). What these factors have in common is that their unit of analysis looks beyond the information-processing perspective of decision making, to where their unit of analysis includes two (if not literally, then in

consideration of social forces – an other entity). Moving beyond the individual as the unit of analysis affords evaluation of processes beyond models of information-processing that better capture outcomes of multi-agent interactions (Cooke et al., 2013; Gorman, 2011). To study constructs such as cooperation and competition, all that is needed is a dynamic interchange between two entities: a dyad (Williams, 2010).

2.4.1 Human-automation cooperation

Existing approaches to human-automation cooperation and resilience consider how automation facilitates collaboration with people in dynamic environments (Allen et al., 1999; Fong et al., 2005; Wagner & Arkin, 2011; Woods et al., 2004; Zieba et al., 2009). Other studies address the effects of socially sensitive automation, such as automation that develops trust through conversational cues, appearance, and behavior (Cassell & Bickmore, 2000; Desteno et al., 2012; Robinette et al., 2013), or automation that engages in good or poor etiquette (Parasuraman & Miller, 2004; Takayama, Groom, & Nass, 2009). Another study on close-proximity human-robot collaboration explored the effects of adaptive automation on human-robot team fluency and subjective ratings of satisfaction, safety, and comfort (Lasota & Shah, 2015). However, these studies either focused on human-automation interaction in terms of reliability and reliance, or were conducted in static task environments.

Few studies have used social exchange theory to explain decisions to cooperate with machine agents in dynamic environments; however, there is potential for such theory to inform the design of future systems. For example, when two or more entities need to coordinate, there is often an interruption. As scholars have observed, interruptions despite their costs to cognitive workload and workflow, often serve a particular purpose – as necessary and timely transfer of information in adaptable and flexible work systems (Grundgeiger & Sanderson, 2009; Rivera-Rodriguez & Karsh, 2010; Walji, Brixey, Johnson-Throop, & Zhang, 2004). However, human-computer interaction studies on interruptions

largely consider the impact on the individual; such as task completion time, error rate, affective factors like annoyance or anxiety (Bailey & Konstan, 2006); the degree of disruption depending on task characteristics (Czerwinski, Cutrell, & Horvitz, 2000); or the role different interruption management strategies – immediate, negotiated, mediated, and scheduled – on task performance (McFarlane, 1999). While these are important efforts toward understanding the costs of interruptions and the effectiveness of different strategies dealing with them, these findings say little about potential tradeoffs of the cost and benefits of interruptions, as would occur in real-world coordination situations. Applying social has the potential to provide such insights because it considers a unit of analysis of at least two, focusing on how features of interactions, such as decisions to interrupt and the tradeoffs that occur, might influence broader system outcomes in an information-imperfect world.

Similar to much of the early interruptions research, other work in human-automation interaction focus on either the information-processing elements, such as physical attributes and ease of use, or organizational level roles of the automation, such as integration with workflow. As a result of this, some approaches turn to designing around people's information-processing limitations (e.g. Bailey & Iqbal, 2008). While these aspects of design are important, and useful in some cases, they provide a limited or static view of the motivations behind people's interactions with automation, how interactions might change as social and work environments evolve, and how such interactions contribute to a system's global goals. Beyond some recent work on human-automation cooperation in relatively static game theoretic situations (e.g. Sandoval, Brandstetter, Obaid, & Bartneck, 2016), there is still little research focusing on interdependent decision-making with more advanced automation and how such decisions lead to value-added cooperation in dynamic coordinative environments.

From a systems perspective, it may be more useful to know how automation behaviors influence people's in-the-moment as well as future decisions to cooperate, and how their joint behaviors relate to macro-level effects and outcomes, such as degree of cooperativeness and overall system performance. Social exchange theory provides a framework for describing the tradeoffs of joint actions, how signals of the actors shape decision-making, and how the social structure of interactions can lead to immediate outcomes and macro equilibriums (Kelley et al., 2003; Schelling, 1973). The current work includes these perspectives but draws from a more specific line of inquiry that posits signals of agent cooperation (rooted in trustworthiness) influences people's decisions to cooperate (Cox, 2004), and that the social structure of the exchange can influence trust and cooperation (Molm et al., 2000).

2.4.2 The role of reciprocity

One of the best-known exchange rules in social exchange theory is reciprocity (Cropanzano & Mitchell, 2005), a mechanism that can promote cooperation. *Reciprocity* is responding in kind, such as repayment of hostility with hostility, or kindness with kindness (Fehr & Gächter, 1998; Sandoval et al., 2016). As a phenomenon reciprocity has accounted for stability and instability in social systems, including the pooling or redistribution of resources within a group to the group's advantage (Ostrom, 2000), the punishing of antisocial behavior despite great costs to oneself (ostensibly for the good of the group) (Gintis, 2000), and as a key mechanism for the enforcement of social norms (Fehr & Gächter, 1998). At its most basic unit, reciprocity can occur in a single exchange. Over time, reciprocity serves as "a mutually gratifying pattern of exchanging goods and services" (Gouldner, 1960, p. 170), and in larger group settings, e.g. societies, may also be reinforced through convention (Young, 1996).

Whereas reciprocity refers to a particular pattern of behavior, cooperation refers to behavior that benefits the group involved. The important fact is that reciprocity occurs through the response of a heterogeneous entity, as a component of interdependent behavior between two or more entities, and is a likely source of conditional cooperation, i.e., *reciprocal cooperation* (Axelrod & Hamilton, 1981). Reciprocity can involve both instrumental or symbolic value, with instrumental value having more concreteness, and symbolic value implying additional value beyond its objective worth, and more likely to vary depending on its source (Cropanzano & Mitchell, 2005). Other facets include positive or negative reciprocity, which refers to altruistic acts or punishment (Fehr & Gächter, 1998), and direct or indirect reciprocity, which respectively correspond to personal enforcement or community enforcement in an exchange with a partner (Nowak & Sigmund, 2005).

To what extent the amount of return in reciprocity is, or should be, to constitute "rough equivalence" (Gouldner, 1960, p. 175) is an open empirical question requiring consideration of how the actors or situation would define equivalence. Reciprocity has been measured in the simultaneous or immediate next response in an exchange, given an initial action (Axelrod & Hamilton, 1981; Barrett, Gaynor, & Henzi, 2002), as well as the overall response value relative to an initial value, within a specified time period (Barrett et al., 2002). Particularly when tied to cooperation, measures of reciprocity look at in-kind responses from the perspective of direct benefit to individuals while considering longer-term outcomes within the relationship or larger community. Because the primary goal of this work is to better understand outcomes of human-automation cooperation for improving outcomes in joint coordination, this dissertation focuses on *positive, instrumental reciprocity* where *reciprocity* describes the degree to which exchange partners return cooperative behavior in-kind, without the partners necessarily ascribing value or exhibiting any defined mental attitude, although values and attitudes may be present.

The following criteria were used to determine if study participants' behavior was reciprocal:

1) If participants' actions return, to the agent's first move, resources of positive value, and
2) The action is costly to the participant in the sense that the amount returned would diminish her ability to maximize her utility if she did not receive the first amount by the agent (similar to Berg et al., 1995; Cox, 2004).

This operationalization of reciprocity considers a limited, shared resource scenario in which a person with altruistic or other-regarding inclinations may return the resource to the agent who, after making the first positive transfer to the participant, now has a lower margin of spare resources than participants. The mere fact that the participant returns the resource to the agent is not evidence of positive reciprocity. A self-regarding person presumably would not return resources to the agent.

Reciprocity also relates to a social exchange conceptualization of trust in that a "trustor" by giving resources is said to place a trust in a "trustee." The trustee then, keeps the trust, i.e., reciprocates, if she returns greater than the value imparted. Trust is therefore defined in terms of the following two actions; first the trustor gives a trustee the right to make a decision, and second, the trustee makes a decision which affects both trustor and trustee (Berg et al., 1995). Since reciprocity is defined as a transactional pattern of interdependent exchanges, inaction would therefore be outside its general realm (Cropanzano & Mitchell, 2005). Although occasionally inaction in cooperative exchange is related to negative reciprocity (Cox, 2004), negative reciprocity has been defined as a separate construct from positive reciprocity rather than as part of the same spectrum, due to the robustness of its effects across situations (Gintis, 2000). Free-loading, a type of behavior that leads to degradation of cooperation, seems to belong to the more general idea of conditional

cooperation (Gächter & Herrmann, 2009) rather than reciprocity specifically. Therefore, not cooperating in the joint task by withholding resources, and the degree to which this leads to degraded cooperation and poor performance, is considered but as a separate construct from reciprocity.

2.4.3 Social exchange structures

Because social exchange occurs within a structure of mutual dependence, in which actors are dependent on one another for valued outcomes, the simplest exchange relation consists of two actors, each of whom controls resources of value for the other (Molm et al., 2000). A defining dimension of social exchange structures is the level of control actors have – in their own outcomes, in their partner's outcomes, or jointly (the extent to which outcomes are controlled by joint actions) (Rusbult & Van Lange, 2003), and *The Atlas of Interpersonal Situations* (Kelley et al., 2003) is a starting guide for identifying the structure of various exchange scenarios, which will not be reviewed here. The main point is that structures are important to consider in understanding outcomes of exchange because they often exert strong effects on behavior, relatively independent of personal goals and motives.

Questions of trust and cooperation are of interest when the two actors have alternative choices within an exchange structure, as illustrated in Figure 1. Compared to negotiated exchanges, reciprocal exchanges were found to produce stronger trust and affective commitment between people (Molm et al., 2000). The difference between a negotiated exchange and a reciprocal exchange is their structure. In negotiated exchange, neither actor can obtain benefit without first agreeing explicitly; the benefits are thus bilateral however unequal they may be. Most economic exchanges fit this category, receiving a service or product for payment, and some social exchanges, like when two parties jointly decide on the division of household chores. In reciprocal exchanges, actors' contributions are separately performed and are not negotiated. Because these choices are made individually, benefits can

be unilateral. These are more common in instances where acts are performed with no guarantee of reciprocity. The risk of incurring a loss is critical to the evolution of cooperation, because it provides exchange partners the opportunity to demonstrate their trustworthiness.

It should be acknowledged that roles within theoretical exchange structures are not often as concrete as their representations might suggest, and decisions often comprise multiple sub-decisions with myriad interacting variables. There are certainly limitations to the extent that real world situations, roles, and decisions can be represented so simplistically. However, including a consideration of social exchange structure in human-automation research can reveal insights currently missing in the knowledge base, including the impact of different structures on cooperation. Such insights can guide information design and decision-making strategies for future human-automation systems.

The following studies presented in this dissertation merge insights from resilience engineering, human-automation interaction research, and a social exchange worldview to better understand cooperation in human-automation coordination. The goal is to better understand potential factors influencing joint performance in a dynamic coordination task. However, designing for and measuring system tradeoffs, such as decisions made when goals potentially conflict, can be difficult to determine in a real world setting with many confounding variables, and where system boundaries are difficult to ascertain.

2.5 Microworlds as a Research Platform

Many important, real-world phenomena such as coordination and cooperation involve people engaging in dynamic, complex decision-making behaviors. Because of this, the precise control and measurement that can be achieved in traditional laboratory studies may generate irrelevant findings. Field studies have many challenges as well, including the challenge of attributing causality due to complexity in the experimental setting and lack of

experimental control. To minimize intrusion in people's work, gathering field data may involve using self-reports and researcher observation. However, self-report often invokes more reflective responses rather than in-the-moment responses, which can differ when studying people's responses to agentic objects (Takayama, 2009). A researcher's presence when observing in person can be intrusive to the task as well, or difficult when the decisions and actions are fast-paced. A solution for this is to use screen capture or video-recorded and coded data; however, video data still for the most part, and what is commercially available, requires manual coding and is thus time-consuming and expensive.

Microworld environments help bridge the gap between the inherent complexity of a field investigation and the control of a laboratory study (Omodei & Wearing, 1995). Building microworlds for research are now achievable within relatively short time frames given the accessibility and abilities of present-day computers and software, though the use of microworlds has long been a tradition in the study of dynamic decision-making. Dynamic decision-making involves a sequence of interdependent, real-time decisions in a changing environment, which microworlds are well-suited to simulate (Gonzalez, Vanyukov, & Martin, 2005). Examples of dynamic decision-making include choosing which routes to take when driving a vehicle, developing and selecting the best strategy while playing a sport, and investing in markets as prices change. In all cases, a sequence of decisions is made in an environment that changes as a function of that sequence, independently of that decision sequence, or both. A microworld can be a direct method for gathering human-agent interaction data while controlling for system level factors that would be difficult to control in the field or difficult to glean from mining naturalistic data. Microworlds are a compromise between experimental control and realism; the assumption is that they embody the essential characteristics of real-world decision environments while providing the experimental control needed to develop explanations of processes, rather than task-specific descriptions of

decision-making, so that results are generalizable across a variety of dynamic decision-making tasks.

2.6 Research Summary and Significance

Existing research in human-automation interaction lacks insight into lateral coordination scenarios with increasingly autonomous agents. Such scenarios demand greater attention to cooperation and the role it plays in people's interactions with automation and subsequent system outcomes. Based on previous work showing people's real tendency to interact with machines in ways that are best explained through social constructs, and considering trends of increasingly autonomous and capable machines, this dissertation aims to provide a starting step for future work on human-automation cooperation. In particular, it integrates perspectives from social exchange theory to explore human-automation interaction dynamics that lead to more resilient systems. At the time of this writing, no human factors studies were found that considered potential social exchange factors and cooperation with, rather than reliance on, automation in a dynamic, joint-task environment.

2.7 Research Objective and Questions

To more clearly understand fundamental mechanisms of cooperation in networked and layered sociotechnical systems, the smallest unit of analysis for cooperation is the focus of this dissertation, the human-agent dyad (Thibaut & Kelley, 1959; Williams, 2010). The goal is to demonstrate that for a system to be resilient, an important complement to people's reliance on automation is their cooperation with automation. A microworld scenario was developed that required human and automated agent dyads to cooperate in a shared-resource task, while varying the cooperativeness of the agent, work environment factors, and the social exchange structure of the joint task. The shared-resource task was developed to ensure people and agents had equal authority and responsibility to avoid potential confounding factors related to differing roles.

Part 1 of this work (Chapter 3) tests the hypothesis that agents' cooperation will affect people's cooperation, leading them to reciprocate with similar resource exchange behaviors. It was thus also expected that these differing levels of agent cooperation would subsequently affect coordination and joint performance. The second hypothesis posits that when people's initial exposure to an agent is in a highly demanding situation, then the effects of agent cooperation will be particularly prominent. People's perception of the agent's intention may be a function of the task environment; therefore, people initially exposed to high-workload and an agent that gives more could be more inclined to cooperate in a subsequent low-workload situation, than people initially exposed to low-workload and an agent that gives less. Part 2 of this work (Chapter 4) asks whether changing the structure of the task, from a negotiated exchange where both partners have input in the decision, to a reciprocal exchange where decisions to provide resources are unilateral and unprompted, will improve cooperation and coordination. This expectation is based on the premise that the symbolic value of unprompted resource provision may induce people's social engagement and prosocial behavior even in risky and high-stress environments.

Chapter 3 Part 1: Effects of Agent Cooperation on Human-Agent Coordination

This section is an adapted version of a manuscript titled "Cooperation in human-agent systems to support resilience: A microworld experiment" that was accepted for publication in *Human Factors: The Journal of the Human Factors and Ergonomics Society* (Chiou & Lee, 2016). Additional analyses were added, mainly in section 3.5.3, to connect this section's findings with Chapter 4 in this dissertation. All headers, formatting, and references were revised to integrate with the rest of this dissertation.

3.1 Background and Motivation

Automation is becoming increasingly autonomous. From self-driving vehicles to sophisticated decision support systems, computational advances have led to a plethora of machines capable of automating whole functions that previously required humans. Emerging new human-automation relationships pressure the existing frameworks for research, design, and evaluation of these joint systems (Woods & Hollnagel, 2005). Humans and machines increasingly enter into coordinative and cooperative relationships, where successful interactions demand teamwork. Resilience Engineering approaches the design of such systems as consisting of interactive agents, collaborating to achieve shared goals in increasingly dynamic and safety-sensitive environments (Hollnagel, Woods, & Leveson, 2006). However, few controlled studies have investigated how resilience depends on automation design, which is the focus of this paper.

For supervisory control, joint performance depends on automation reliability, the ability to assess automation reliability, and appropriately relying on or complying with automation (Parasuraman & Riley, 1997). However, with more advanced automation, reliability may be less critical than the resilience of the system (Zieba et al., 2010). Resilience refers to the ability to manage sustained adaptability of a layered network system (Woods, 2015). This differs from resilience as rebound, to restore a system to previous conditions

before disruption, or resilience as robustness, having an expanded set of models to respond to disturbances (Woods, 2015). Rather, resilience as sustained adaptability considers how interacting agents cooperate, drawing on shared resources to accommodate surprises. Designing for resilience suggests collaborative or cooperative control might be an important complement to supervisory control.

For human-automation interaction, collaborative control allows both human and automated agents to combine competencies and negotiate conflicts in dynamic settings (Zieba et al., 2010). Collaborative control supports resilience through adaptation to novel situations, accommodation of bugs in the system itself, and recovery from errors in following procedures (Woods et al., 1994). Furthermore, resilience calls for the continual renewal of shared goals, reciprocity, and the willingness to accommodate others as unexpected demands require shifts of individual and shared priorities (Woods, 2004). System performance is less about how well predetermined priorities are reliably executed, and more about how well goals and actions are adapted for the greater good of the system. Thus, rather than collaborative control, which often refers to managing functional dependencies by adapting levels of autonomy, cooperative control may be more appropriate for human-automation scenarios where cooperation precedes coordination.

Resilience in uncertain environments with complex interdependencies demands alternatives to the more explicit supervisor-subordinate procedures. These alternatives include informal interactions that promote coordination and cooperation (Rochlin et al., 1987). Coordination is broadly defined as dependency management (Malone & Crowston, 1994), and concerns task scheduling and assignment, whereas cooperation is identified in actions that benefit a group but impose individual costs (adapted from Dugatkin, Mesterton-Gibbons, & Houston, 1992). The need to compromise individual goals for a shared goal can arise from what Clark (1996) calls private and public goals, Ostrom (2000) calls self and

collective interest, Hoc (2001) calls ego-centric and collaborative interests, and Woods (2004) calls local and global goals. These constructs share the idea that cooperation emerges through reconciling competing goals with the help of social processes. When interdependent activity involves shared resources and conflicting goals, cooperation is essential for coordination.

The importance of cooperation in resilience is highlighted in the margin of maneuver—a cache of actions and resources that allows the system to function despite unexpected perturbations in the work environment (Woods & Branlat, 2010). Systems without adequate margins become brittle, unable to withstand unexpected demands. Based on studies of emergency department interactions with other hospital units, Stephens, Woods, Branlat, and Wears (2011) identify three classes of strategies to create or maintain a margin of maneuver: defensive strategies, autonomous strategies, and cooperative strategies. Defensive strategies increase the margin of a unit by restricting another unit. Autonomous strategies involve local focus, such as reducing interactions with other units because the benefits of sharing resources are not believed to be worth the costs. Cooperative strategies involve the effort of two or more units that, through coordinated and collective action, recognize or create common-pool resources from which both units can draw. Such actions help avoid the tragedy of the commons when resources are limited (Ostrom, 1999), and may enhance system resilience by giving the overall network a larger pool of resources to draw upon for greater margin of maneuver.

In light of these properties of resilience in volatile environments and increasingly capable automation, understanding what contributes to human-agent cooperation seems crucial. When problems cascade, demands for cognitive activity increase as do the demands for coordination across agents (Woods & Patterson, 2001). These increased demands may accentuate bonding or fracturing of teams depending on the nature of their interdependence

and the information available regarding the trustworthiness of agents (Gao et al., 2006). Cooperative strategies require assuming initial generosity (Axelrod & Hamilton, 1981), trust that the other party will reciprocate and not exploit shared resources, and an understanding that the benefits of cooperation will outweigh the costs. Without such trust between agents, an agent may resort to defensive strategies to preserve individual gain at the cost of another (De Dreu et al., 2010).

A substitute for formal controls, trust guides human behavior toward automation when a complete understanding is difficult or impractical (Lee & See, 2004). Such uncertainty is increasingly common, due to technological advances and the environments in which automation are deployed. In high-stress environments, human agents may be even less likely to proactively interact with automation, particularly when experiencing physical and cognitive fatigue (Casper & Murphy, 2003). Although trust in automation has been said to be asymmetrical, in that automation does not trust back (Lee & See, 2004), recent work with more autonomous automation (e.g., agents or robots), suggests a more symmetrical view and social exchange situations may apply to the concerns raised by resilience engineering (Bray et al., 2012; Fink & Weyer, 2014; Wagner & Arkin, 2011). In cooperative exchange, people often choose partners depending on the instrumental value of the exchange, even though people's trust, affective regard, and sense of solidarity with exchange partners are strongly influenced by the symbolic act of reciprocity (Molm, Schaefer, & Collett, 2007). Reciprocity and trust influence cooperative behavior, especially when novel or complex situations involve risk or uncertainty (Delton et al., 2011; Molm et al., 2000; Riedl et al., 2011).

Existing approaches to human-automation cooperation and resilience consider how automation can facilitate collaboration with people in dynamic environments (Allen et al., 1999; Fong et al., 2005; Wagner & Arkin, 2011; Woods et al., 2004; Zieba et al., 2009). Other studies address the effects of socially sensitive automation, such as developing trust

through conversational cues, appearance, and behavior (Cassell & Bickmore, 2000; Desteno et al., 2012; Robinette et al., 2013), or engaging in good or poor etiquette (Parasuraman & Miller, 2004; Takayama et al., 2009). Another study explored the effects of adaptive automation on human-robot team fluency and subjective ratings of satisfaction, safety, and comfort (Lasota & Shah, 2015). However, these studies focused on reliability and reliance, or were conducted in static task environments. At the time of this writing, we were unable to find studies that considered cooperation with automation in a dynamic, shared-resource task environment.

As a first step in understanding cooperation in networked and layered sociotechnical systems we consider the human-agent dyad (Thibaut & Kelley, 1959; Williams, 2010). Our goal is to demonstrate that for resilient systems, an important complement to reliance on automation is cooperation. To achieve this goal, we developed a microworld that required human and automated agent dyads to cooperate on a dynamic, shared resource task. We present two experiments; Experiment 1 tests the hypothesis that an agent's cooperation will affect people's cooperation, leading them to reciprocate with similar resource exchange behaviors. Experiment 2 tests the hypothesis that when people's initial exposure to an agent is in a highly demanding situation, then the effects of agent cooperation will be particularly prominent.

3.2 Overview of Scheduling Task and Microworld Environment

A microworld hospital scheduling scenario was developed in Java and XML with the Android SDK, to assess cooperative behavior in a joint human-agent task with shared resources. A 15" laptop computer running Genymotion and a standard computer mouse were used. Participants acted as hospital schedulers, whose task involved assigning patients and staff to hospital rooms, and coordinating shared staff resources with a neighboring hospital managed by an automated agent. Interactions with the agent were bilateral and limited to

requesting staff and responding to requests for staff. The study tested different levels of agent cooperation on participants' behavior and joint performance. In each of their four experimental trials, 36 participants in Experiment 1 experienced a slow-tempo period followed by a fast-tempo period (the slow-to-fast group), and 36 participants in Experiment 2 experienced a fast-tempo period followed by a slow-tempo period (the fast-to-slow group). Testing these two tempo sequences allowed us to assess the effects of agent cooperation level in the context of recovery following a perturbation, and across variation in work environments.

3.3 Experimental Design

3.3.1 Study participants

Participants were recruited near a Midwestern university through flyers and online postings and received 10 dollars at the end of the hour-long study. Self-reports showed age ranged from 18 to 56 with a mean age of 22 in Experiment 1 and a mean age of 24 in Experiment 2; gender was roughly 50% female and 50% male; 97% of participants used the computer daily; 47% in Experiment 1 and 64% in Experiment 2 reported using the computer for playing games.

3.3.2 Independent variables

Each experiment tested "agent cooperation level" as a within-subjects variable, meaning each participant was exposed to a high-cooperation and a low-cooperation agent (Figure 2). Agent cooperation levels were operationalized in the agents' resource-requesting and resource-sharing behaviors. Exposure to the agents was counterbalanced, so participants worked with one agent in their first two trials, and the other agent in their final two trials. Within each trial there was one slow-tempo period and one fast-tempo period, and "tempo" was tested as a within-subject variable. "Scheduler" was a within-subject variable, to distinguish between the participant and the automated agent. Therefore, each study was a

mixed within and between 2 x 2 x 2 x 2 design that tested agent cooperation level (high or low cooperation), tempo (fast-tempo period or slow-tempo period), and scheduler (person or agent) as within-subject variables, and cooperation order as a between-subject variable (high-cooperation followed by low-cooperation, or low-cooperation followed by high-cooperation).

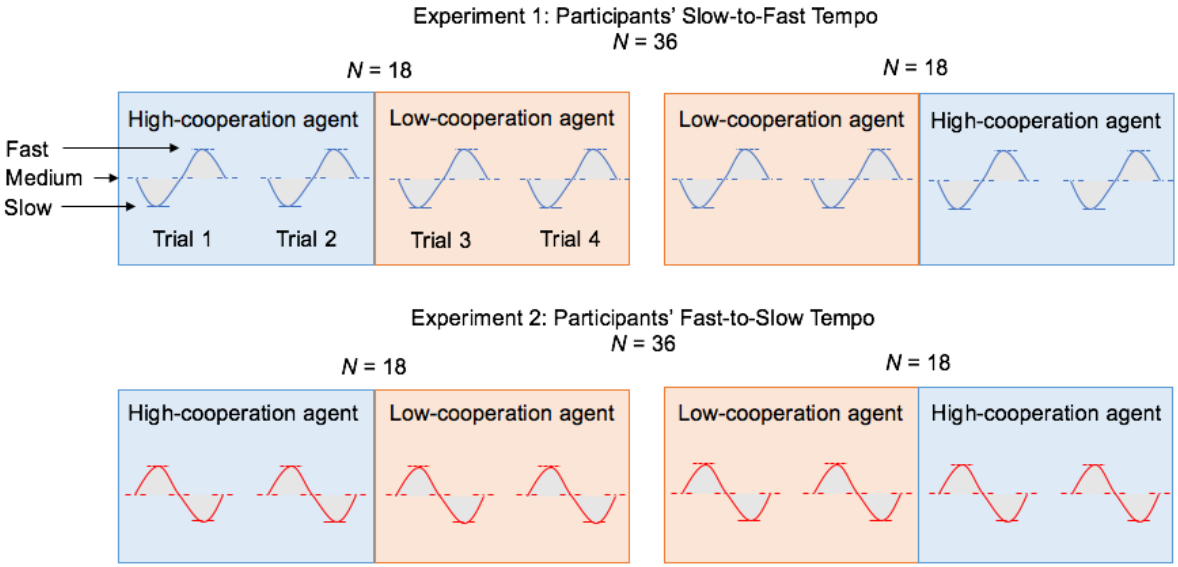

Figure 2. In Experiment 1, 18 participants experienced the high-cooperation agent then the low-cooperation agent, in a slow-to-fast tempo sequence. The other 18 participants experienced the counterbalanced order of cooperation agents. Experiment 2 participants experienced the fast-to-slow tempo sequence.

3.3.3 Tempo levels – slow-tempo and fast-tempo periods

Each 8-minute trial was divided into five 96-second intervals, representing a slow, medium, or fast rate at which patients entered the waiting rooms. In the slow-to-fast sequence of Experiment 1, participants experienced the interval order: medium, slow, medium, fast, and medium. In the fast-to-slow sequence of Experiment 2, participants experienced the interval order: medium, fast, medium, slow, and medium (see Figure 2). For easier communication however, "tempo period" refers to the half of the trial that includes the named interval, e.g. the "slow-tempo period" in the slow-to-fast sequence was one medium interval, one slow interval, and half a medium interval. The slow interval had two patients enter the waiting room, evenly distributed throughout that interval, whereas the medium

interval had five patients, and the fast interval had eight patients. The medium-tempo interval was determined in pilot testing as an engaging pace, without high pressure.

In both experiments, the agent's tempo sequence was always the opposite of the participant's tempo sequence. For example, in participants' slow-to-fast sequence of Experiment 1, agents would experience the fast-to-slow sequence. The complementary sequences allowed for an obvious coordination solution if the dyad cooperated and shared staff, making it possible to assess how differences in cooperative behavior contributed to breakdowns or delays in resource sharing. Participants (or agents) in the slow-tempo period would more likely have excess staff, whereas participants (or agents) in the fast-tempo period would need additional staff to meet demand in their hospital.

Experiment 1 was thus designed to investigate how participants reacted to an agent that requested staff more often and shared staff less often during participants' slow-tempo period, and an agent that requested staff less often and shared staff more often, as participants moved into a fast-tempo period. Experiment 2 was designed to investigate how an initial perturbation, or fast-tempo period, might differently affect participant and joint cooperation. Taken together, both experiments allowed for comparison of cooperative behavior prior to a fast-tempo (Experiment 1) and following a fast-tempo (Experiment 2) to assess team cooperation and ability to adapt to environment changes in a timely manner.

3.3.4 Agent cooperation levels – high-cooperation and low-cooperation agents

Agent cooperation was operationalized as high-cooperation or low-cooperation depending on how much its behaviors emphasized joint outcome or individual outcome, respectively. Table 1 summarizes the resource-sharing and resource-requesting behaviors for the agents.

Table 1. Agent resource-sharing and requesting behavior by cooperation level

Agent Cooperation Level	Resource-Sharing Behavior (Acceptance Rate)	Resource-Requesting Behavior
High cooperation	100 % with 1-2 patients 75 % with 3-4 patients 50 % with 5-6 patients	1. Checks if it needs a resource 2. Checks if the participant has this resource available 3. Requests the resource
Low cooperation	50 % with 1-2 patients 25 % with 3-4 patients 0 % with 5-6 patients	1. Checks if it needs a resource 2. Requests the resource

Note. The agents' acceptance rates are keyed to the number of patients in the agent's waiting room.

The high-cooperation agent provided resources at a higher rate than the low-cooperation agent, depending on its queue length. When a participant requested an available resource, and the high-cooperation agent had two or fewer patients in its waiting room, the agent would accept 100% of the time, 75% of the time with three to four patients, and 50% of the time with five to six patients. The low-cooperation agent provided resources at 50% less than the high-cooperation agent in each of the waiting room conditions, or 50%, 25% and 0% respectively. The low-cooperation agent thus expressed less individual risk and prioritized its own performance, compared to the high-cooperation agent. Note that the low-cooperation agent is not competitive, which could involve requesting all staff resources and refusing to return them; it is still cooperating in this context by sharing and requesting staff, just at a lower level. These conceptualizations of the agents' behaviors are supported by the conceptualization of cooperative behavior in social exchange as actions that 1) gives a positive value to another entity that is 2) risky for the giver in the sense that would lead to a loss if none were returned by the receiver (Cox, 2004).

Cooperation was also expressed through requesting behavior. The high-cooperation agent would check if it needed a resource, then check if the participant had the resource before requesting. The low-cooperation agent only considered its own need for staff, which made it more likely to make requests insensitive to the participants' needs. Agents were

programmed to have an 8-second and 2-second delay when assigning patient or staff and when collecting staff, respectively, to simulate the pace of a human player, established during pilot testing. This delay also avoided continuous interruption of the participant and allowed a window for participants to request unassigned agent staff.

3.3.5 Dependent variables

Because we wanted to know how an agent's cooperative behaviors would influence a person's cooperative behaviors, the dependent variables in this study were participants' resource-requesting and resource-sharing behaviors. These were measured as the number of staff they requested and the number of agents' requests they accepted. Agents' behaviors were tallied independently. Individual performance was measured as the sum of patients treated in a trial, and joint performance was the sum of the patients participants and agents treated in a trial.

3.4 Procedure

Participants were introduced to the microworld scheduling task and interface. Participants were told they could request staff at any time from an automated agent, scheduling its own hospital in the background. To avoid guiding participants to use a particular strategy, participants were told to, "treat as many patients as possible." If participants asked if this included cooperating with the agent, the goal was restated and they were told it was up to them how to reach the goal. No information was given about the neighboring hospital agent's behavior or potential changes in tempo.

Participants were then exposed to two, 2-minute practice trials at a medium-tempo. At the end of each trial, including experimental trials, a bar graph displayed the patients they treated in the context of patients treated in both hospitals. The graph emphasized the goal of maximizing a joint score and acted as motivational performance feedback.

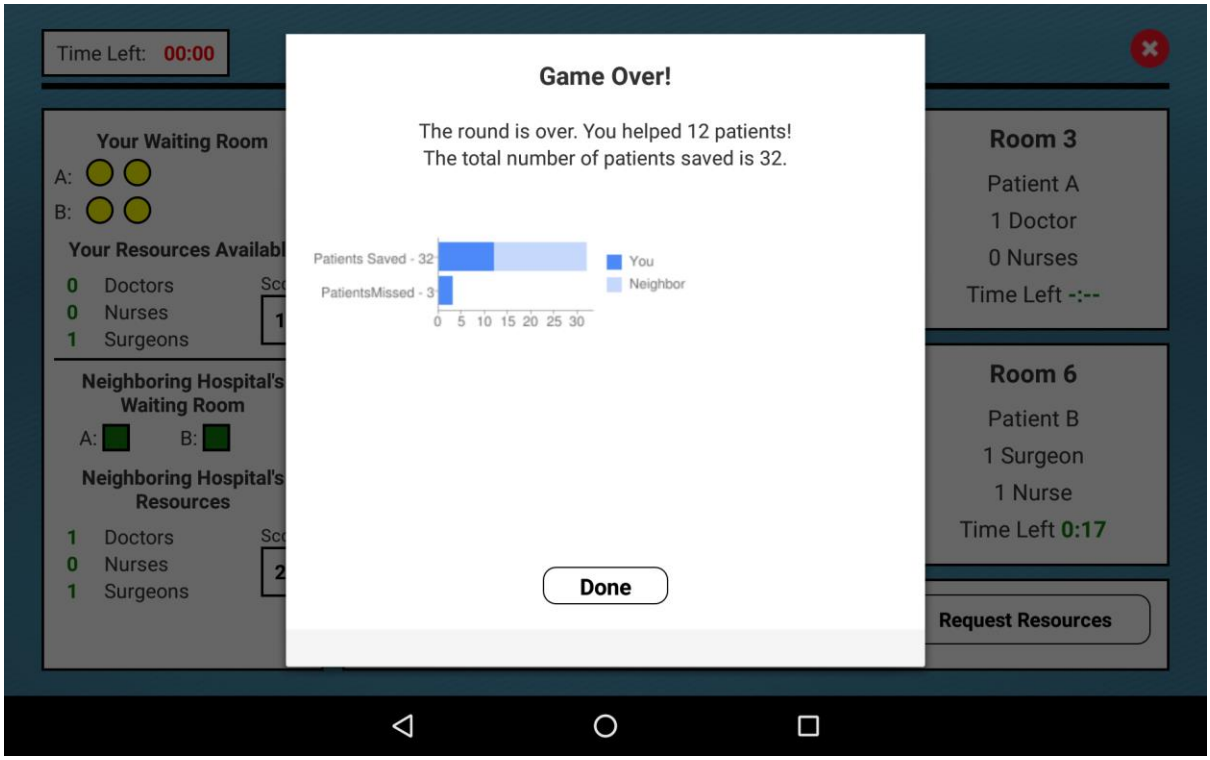

Figure 3. Feedback provided to participants showed a split bar of patients treated in each hospital, with a darker shade representing the participant's contribution. The sum of patients missed was also reported as a split bar when applicable.

At the beginning of a trial, participants and their agent partner were each given three nurses, two doctors, and two surgeons, along with one patient in their waiting room queue (Figure 4). Hospital staff were the only shareable resources. There were two types of patients that entered the queue with randomized 50% chance; patients "A" required a doctor and a nurse, patients "B" required a surgeon and nurse. Both types of patients therefore required a nurse, and to use all of one hospital's rooms a scheduler would need to obtain all nurses from both hospitals, and some additional doctors or surgeons.

The main interface components comprised a bottom control panel with three button options, Assign Patient to Room, Assign Resource to Room, or Request Resources (Figure 4). At the center of the interface were the six hospital rooms, also buttons. To help with resource management, a side panel displayed available staff, patient waiting rooms, and a "score" for each hospital. Color-codes were used to signal waiting room queue length status: green for one to two, yellow for three to four, and red for five to six patients. Six patients was

a waiting room's maximum capacity, and participants were told additional patients scheduled to arrive would be turned away. Participants could see the exact number of patients in their waiting room, with each circle representing one patient (Figure 4). However, only the color status of the agent's waiting room was available to participants. This decision was inspired by Dabbish and Kraut (2004), which showed summarized displays can be more effective than detailed displays for communicating status information in a cooperative task. A summarized display also contributed to the uncertainty of interacting with agents, allowing a focus on the social-affective aspects of participants' decision-making rather than the information-processing aspects. Two squares in the agent's waiting room showed it also received A and B patients, and a timer at the top of the screen counted down to the end of a trial.

Figure 4. A screenshot of the interface at the beginning of a trial shows a bottom control panel, left side panel, six hospital rooms, and a timer.

Table 2 describes the main actions of the scheduling task from the participant's point of view. Level 1 refers to the bottom control panel options in Figure 4, with subsequent navigation options as the following levels.

Table 2. Participants' main actions and navigation pathways in negotiated exchange

Level 1	Level 2	Level 3		Level 4
Actions and Pathways Related to Scheduling				
Assign patient to room	*Patient A* *Patient B*	*(Select available room)*		--
Assign resource to room	*Doctor* *Nurse* *Surgeon*	*(Select available room)*		--
Actions and Pathways Related to Resource-Exchanging				
Request resource	*Doctor* *Nurse* *Surgeon*	"The other hospital has given you a (doctor/ nurse/surgeon)"		--
		"Your request has been denied by the other hospital."		
"The neighboring hospital requests your assistance! They would like to have a (doctor/nurse/surgeon)."	*Accept request*	"Send (doctor/nurse/ surgeon) to neighboring hospital?"	*No*	--
			Confirm	"A (doctor/nurse/ surgeon) has been sent to the other hospital."
	Deny request	--		--

Note. Level 1 indicates the first level of options and the subsequent levels describe potential pathways for the options selected. Italicized texts refer to button options and quotations refer to pop-up communications. Two dashes indicate levels not applicable to that pathway.

Interface highlighting and feedback aided participants in the microworld. For example, when assigning patients and staff, rooms currently treating patients would be greyed out. Once a patient and staff were assigned to a room, "patient treatment" began automatically, lasting 60 seconds. After treatment, participants needed to click a "Collect Resources" button that appeared on top of the respective room to free the room and staff for reassignment (action not included in Table 2). The system generally did not allow erroneous assignments; however, if participants attempted to take action using unavailable patients or staff, floating text would appear with the error and last several seconds before fading.

To request staff, participants could select "Request Resource" from the bottom control panel, then select nurse, doctor, or surgeon. Immediately following this, floating text

informed participants if their request was accepted or denied (Table 2, Level 3). If accepted, the resource would transfer from the agent's hospital to participants' hospital and the staff numbers in both hospitals would update. Incoming requests to participants blocked the control panel, forcing an interruption (Figure 5). Staff that transferred would remain at the hospital unless requested and accepted back.

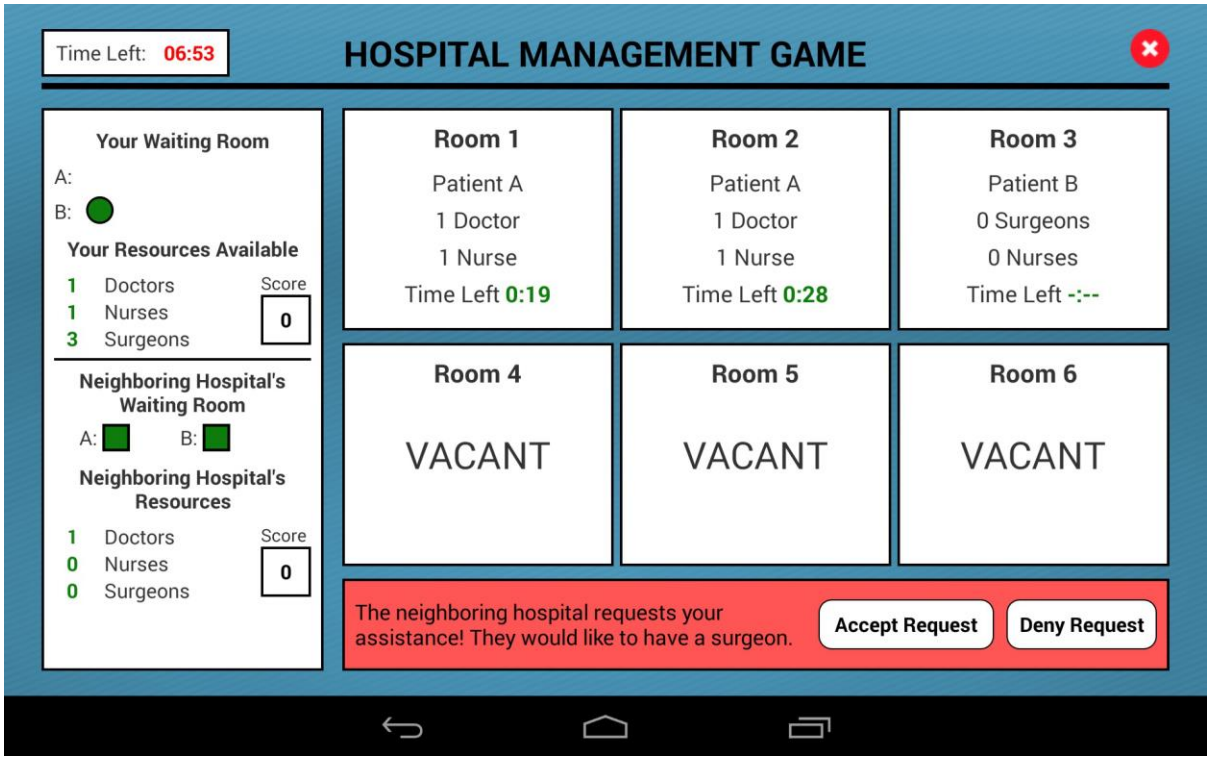

Figure 5. A screenshot of the microworld environment interface that shows the bottom control panel obscured by a resource request.

3.5 Results and Discussion

In Experiment 1, where participants experienced a slow-tempo period followed by a fast-tempo period, we assess whether the high- and low- cooperation behaviors affected participants' behaviors using ANOVA with three within-subjects variables (scheduler, cooperation level, tempo) and one between-subjects variable (cooperation order). Participants' and agents' requesting and accepting behavior are discussed in the context of these resource-sharing variables, and mean performance (number of patients treated, individually and jointly) are considered in the context of agent cooperation. To assess the

validity of averaging performance across trials, participants' mean performance across trials was compared. No significant difference was found, suggesting minimal learning effects.

In Experiment 2, a different group of participants experienced a fast-tempo period followed by a slow-tempo period. Resource-sharing in Experiment 2 is compared to Experiment 1 by how people's cooperation differed following a perturbation, an initial fast-tempo period. We follow the same assessments as in Experiment 1 and supplement our explanations with additional visualizations and analyses. Analyses were conducted using the 'stats' package in R (R Core Team, 2014); data frame manipulations and descriptive statistics were conducted using the 'dplyr' package (Wickham & Francois, 2015); data visualizations and figures were created using 'ggplot2' (Wickham, 2009). All figures are labeled from the participants' perspective.

3.5.1 Experiment 1: Slow-tempo period followed by a fast-tempo period.

Figure 6 shows agents' and participants' mean requests in the background narrow bars, overlaid by their mean acceptances, the foreground wider bars. Data are faceted by cooperation order (rows), cooperation level (columns), and tempo period (columns within cells). There were more requests in the fast-tempo periods (quadrants' right columns, narrow bars) compared to the slow-tempo periods (quadrants' left columns, narrow bars), demonstrating that agents generally behaved as designed within the microworld, and that participants understood the joint task and engaged in requesting staff accordingly ($F(1, 34) = 91.89$, $p < 0.01$). Participants' increased requests moving from slow-tempo to fast-tempo shows the effect of tempo ($F(1, 34) = 47.47$, $p < 0.01$), and that people engaged agents as they needed more staff. Mean acceptances (Figure 6 wide bars) were not significantly different overall between tempo periods (quadrants' left and right columns) ($F(1, 34) = 0.48$, $p = 0.5$) due to the complementary tempo design and joint structure of the task. However, acceptances across cooperation levels did interact with tempo; participants' mean

acceptances were greater in their slow-tempo period compared to their fast-tempo period (dark wide bars, quadrants' left columns compared to right columns, $F(1, 34 = 96.58, p < 0.01)$. Furthermore, when participants were supposed to be accepting requests during their slow-tempo (quadrants' left wide bars), their mean acceptances were fewer than agents' mean acceptances during its slow-tempo (quadrants' right wide bars) in three of the four comparisons. The same was true in participants' fast-tempo compared to agents' fast-tempo (participants' slow-tempo), participants' mean acceptances were fewer than the agents' in three of the four comparisons. In addition, mean acceptances were generally higher in the bottom row of Figure 6, particularly in the high-cooperation condition, which reflects an interaction between cooperation order and cooperation level ($F(1, 34) = 4.31, p = 0.05$).

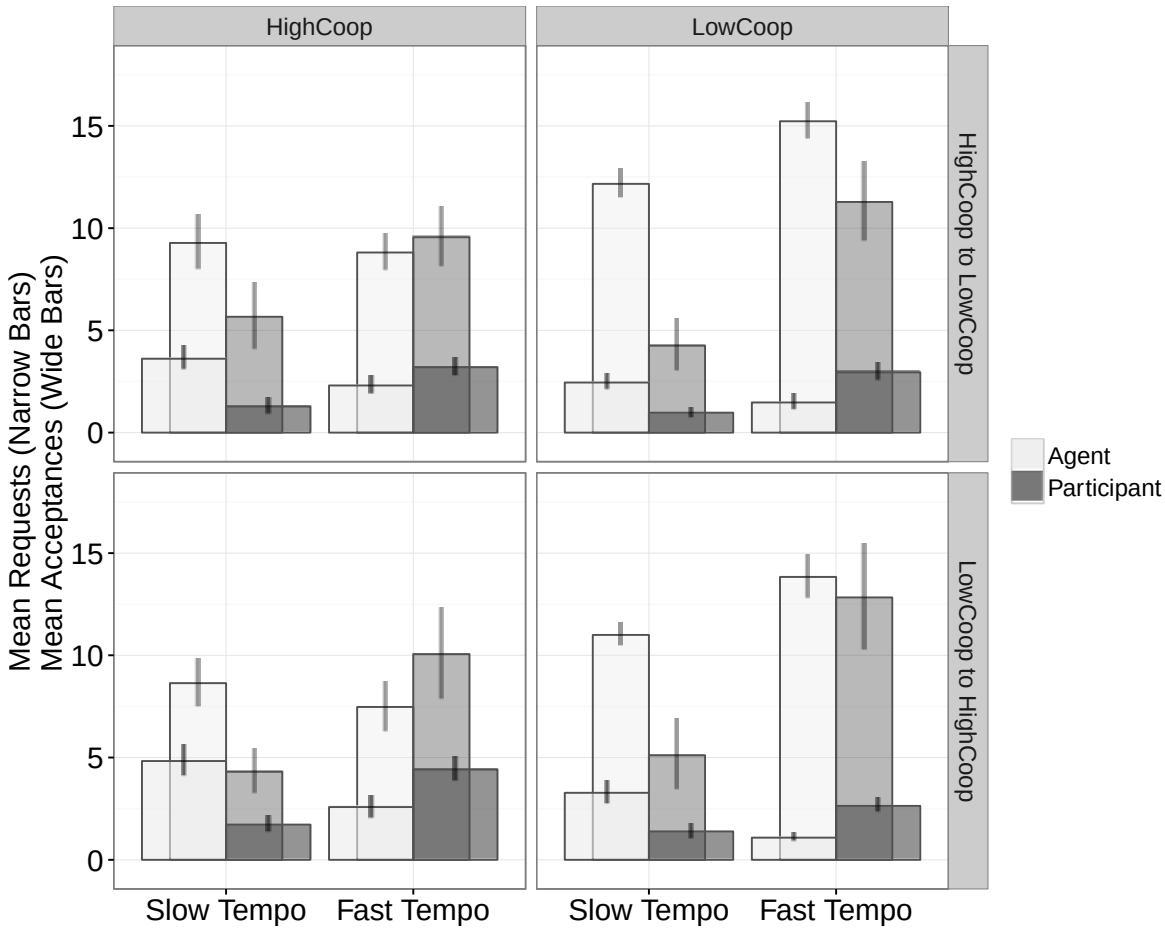

Figure 6. Agents' (bars on the left) and participants' (bars on the right) mean requests (narrow bars) and mean acceptances (wide bars) with 95% CIs (confidence intervals).

Appropriate cooperation means participants should generally accept more requests during their slow-tempo period, and make more requests during their fast-tempo period. The question is if agent cooperation differentially affected these behaviors. While agents' mean requests were greater than participants' mean requests across conditions ($F(1, 34) = 26.46$, $p < 0.01$), and agents' mean acceptances were greater than participants' mean acceptances across conditions ($F(1, 34) = 24.37$, $p < 0.01$), overall mean requests in the high-cooperation condition were fewer compared to the low-cooperation condition (Figure 6 comparing the total narrow bars in the left column with the total narrow bars in the right column) ($F(1, 34) = 35.62$, $p < 0.01$). Across conditions, the low-cooperation agents' mean requests were greater than participants' mean requests, reflecting the significant interaction for scheduler and cooperation level ($F(1, 34) = 13.16$, $p < 0.01$). In sum, participants requested more from the low-cooperation agent than from the high-cooperation agent.

Mean acceptances, in contrast, were greater in the high-cooperation condition compared to in the low-cooperation condition (Figure 6, comparing wide bars in the left column and right column) ($F(1, 34) = 37.67$, $p < 0.01$). Thus, participants contributed to more productive interactions, that led to an exchange of staff, in the high-cooperation condition compared to the low-cooperation condition. Participants' mean acceptances were fewer than agents' mean acceptances, and even fewer in the low-cooperation condition compared to the high-cooperation condition ($F(1, 34) = 21.33$, $p < 0.01$). Despite interacting more often with the low-cooperation agent, indicated by the greater number of requests, exchanges were less productive – less likely to lead to resource exchange. This shows participants were still willing to engage the low-cooperation agent, making more requests like the low-cooperation agent, rather than saving effort or adopting a more autonomous strategy. It also means participants as a group did not exploit the high-cooperation agent's relative generosity by requesting more often and refusing to return staff when requested.

The dyad handled resource exchanges differently in high-cooperation versus low-cooperation conditions, particularly in their requests. In Figure 6 (left column, narrow bars), agents' mean requests in participants' slow-tempo were greater than participants', and this reverses when participants reached their fast-tempo – their mean requests are greater than the agents'. In the low-cooperation condition (right column, narrow bars), agents' mean requests are greater than participants' in their slow-tempo, although this does not increase participants' mean acceptances (wide bars). This may have led to agents' subsequent need for staff and the mutually increased requests in participants' fast-tempo. These observations are supported by the significant interaction for cooperation level and tempo ($F(1, 34) = 21.14$, $p < 0.01$).

To determine discretionary cooperation separate from task-induced cooperation, the number of valid requests (requests for staff that were available) and the number of valid acceptances (acceptances made when possessing the available resource) were calculated, regardless of whether or not it was appropriate to request or accept. Comparing participants' discretionary cooperation against agents' discretionary cooperation, participants accepted on average 35.8% (*SD* = 18.83%) of the high-cooperation agent's requests in a trial, whereas the high-cooperation agent accepted on average 54% (*SD* = 20.82) of participants' requests in a trial. The average ratio of percent acceptances per trial with the high-cooperation agent was 71% (*SD* = 32.33%); in other words, participants' returned 71% of the high-cooperation agent's discretionary cooperation. Compared to the low cooperation agent, which accepted 28% (*SD* = 13.49%) of participants' requests, participants accepted 23% (*SD* = 12.8%) of the low cooperation agents' requests. Participants returned 95% (*SD* = 69.63%) of the low-cooperation agents' discretionary cooperation. This large proportion mostly reflects that both participants and the low-cooperation agent accepted a similar number of requests from one another, which was in general a very low amount.

Did more productive interactions with the high cooperation agent lead to higher mean performance? Figure 7 shows mean individual performance, the sum of patients treated by each scheduler, and mean joint performance, the sum of participants' and agents' performance. Because of task interdependence – staff unused by participants were often used by the agents and visa versa – the difference between cooperation conditions is small. In particular, the low-cooperation agents' higher mean performance compensates for the participants' lower mean performance, and reflects the low-cooperation agent's more autonomous strategy favoring individual team member performance. However, the lower mean joint performance supports the observation that the low-cooperation agent's behavior led to less productive participant behaviors, with the high-cooperation condition producing higher joint performance (Figure 7) ($F(1, 34) = 7.37$, $p = 0.01$). When participants transitioned from the low-cooperation to the high-cooperation agent, the difference between the high-cooperation and low-cooperation conditions appears larger (Figure 7, right column differences are larger compared to the left column differences); however, cooperation order and cooperation level did not produce a significant interaction. These findings for joint performance suggest that an inconsiderate interrupter that requested staff indiscriminately, and was less generous in providing staff – the low-cooperation agent's persona – could lead to lower joint performance in a task involving people.

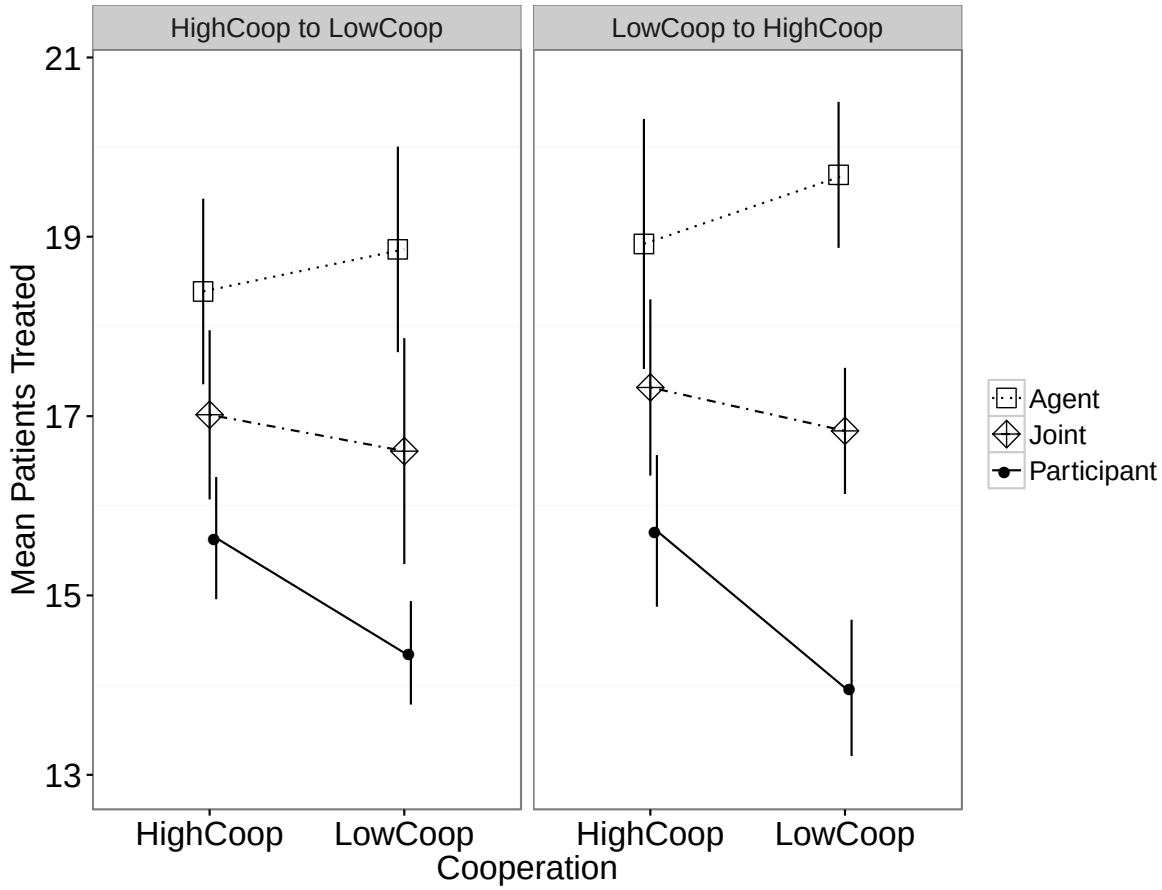

Figure 7. Mean patients treated with 95% CIs show individual performance and joint performance, with joint performance lower in the low-cooperation condition compared to the high cooperation condition. Joint performance is plotted as the average of participants' scores and agents' scores.

While high-cooperation was defined as being considerate in the timing of requesting resources, and providing resources generously at the right time, results show fewer requests occurred with the high-cooperation agent compared to the low-cooperation agent. However, both agent and participant behaviors contributed to the higher joint performance in the high-cooperation condition, given the appropriate timing of their actions. Overall, these results show the value of considering responsive (responding to requests) and proactive (making requests) behaviors as part of evaluating appropriate cooperation in a joint task, and that social exchange factors may be relevant for human-agent interaction – agent cooperation can affect human cooperation.

3.5.2 Experiment 2: Fast-tempo period followed by a slow-tempo period

Requesting behavior in Experiment 2 was similar to Experiment 1; agents' mean requests were greater than participants' mean requests ($F(1, 34) = 124.87$, $p < 0.01$) and mean requests were greater in the low-cooperation condition compared to the high-cooperation condition ($F(1, 34) = 93.72$, $p < 0.01$). In addition, mean acceptances were greater in the high-cooperation condition compared to the low-cooperation condition ($F(1, 34) = 33.27$, $p < 0.01$), and the two scheduler's mean requests were greater in their respective fast-tempo periods, demonstrating the agents behaved as designed and that participants were engaged in the microworld task ($F(1, 34) = 169.43$, $p < 0.01$). Furthermore, participants and agents differed in their requests, particularly depending on cooperation levels; the low-cooperation agents' mean requests were greater than participants' mean requests across conditions ($F(1, 34) = 89.85$, $p < 0.01$).

Contrary to expectations, mean requests in the slow-tempo period were greater than in the fast-tempo period ($F(1, 34) = 223.70$, $p < 0.01$). This highlights the first difference between Experiment 2 and Experiment 1. Figure 8 (quadrants' right columns, narrow bars) partially explains these findings; agents' mean requests during participants' slow-tempo were greater relative to the other conditions and relative to Experiment 1, with an especially large portion of requests made by the low-cooperation agent ($F(1, 34) = 41.92$, $p < 0.01$). This is also reflected in the significant interaction term for scheduler and tempo, where the agents' mean requests were greater that participants' mean requests during participants' slow-tempo period ($F(1, 34) = 222.18$, $p < 0.01$); the significant interaction term for cooperation level and tempo, where mean requests were greatest overall in the low-cooperation, slow-tempo period ($F(1, 34) = 41.92$, $p < 0.01$); and the significant interaction term for scheduler, cooperation level, and tempo ($F(1, 34) = 20.48$, $p < 0.01$).

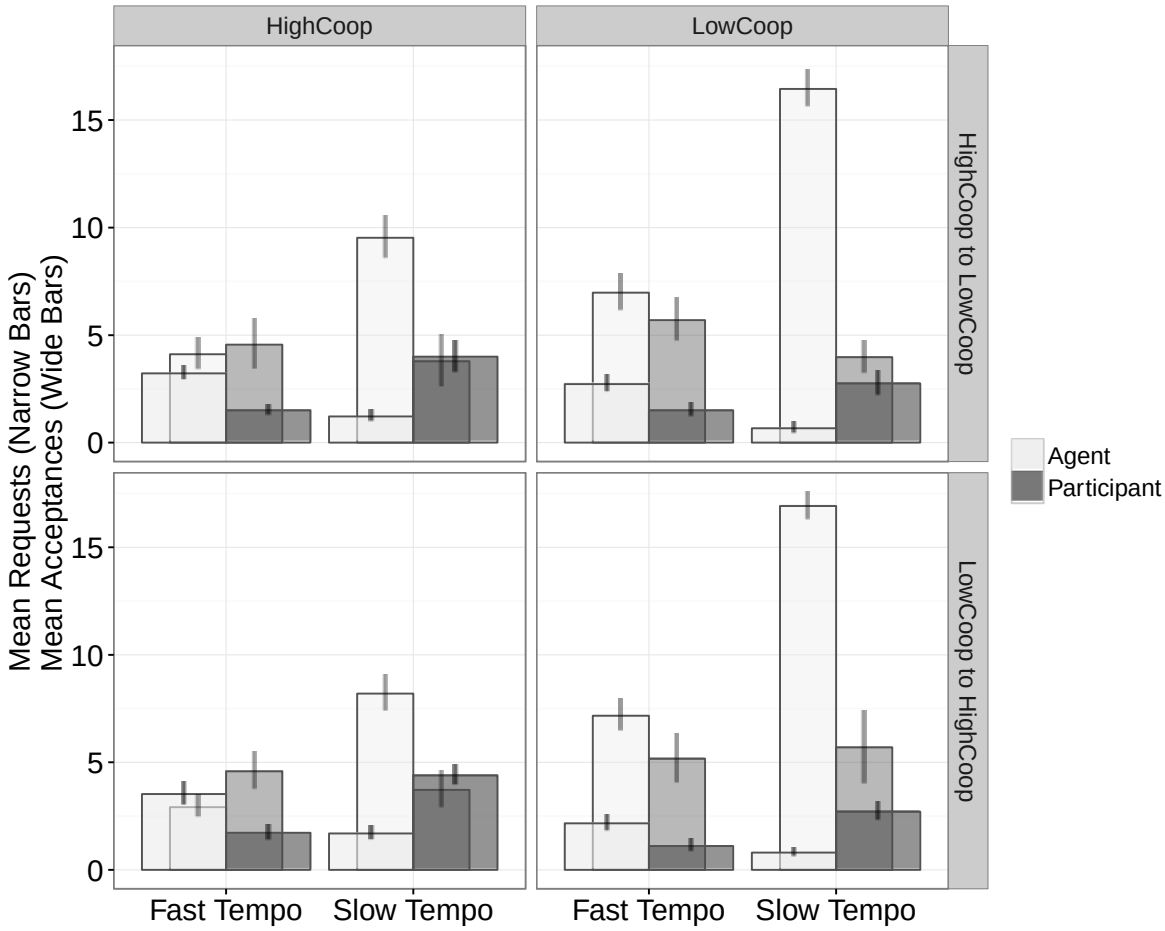

Figure 8. Agents' (bars on the left) and participants' (bars on the right) mean requests (narrow bars) and mean acceptances (wide bars) with 95% CIs (confidence intervals).

Because agent actions were partially contingent on participants' actions, a complementary explanation for the higher number of requests in the slow-tempo period is that participants delayed and did not request staff early, when needed, during their fast-tempo period. This delay caused a backlog of patients in their hospitals, which reverberated through their requesting and accepting behaviors in the second half of the trial. The greater number of agent requests are thus likely due to participants retaining staff leading into the agents' fast-tempo period (and participant's slow-tempo period), resulting in agents running out of staff more quickly compared to Experiment 1. This failure of participants to quickly adapt in the early fast-tempo might be described as a failure to maintain their margin of maneuver (Stephens et al., 2011), which led to subsequent system breakdowns and delays in the joint

task. A visualization of the request sequence shows that participants' requests peaked at the end of their fast-tempo period (Figure 9).

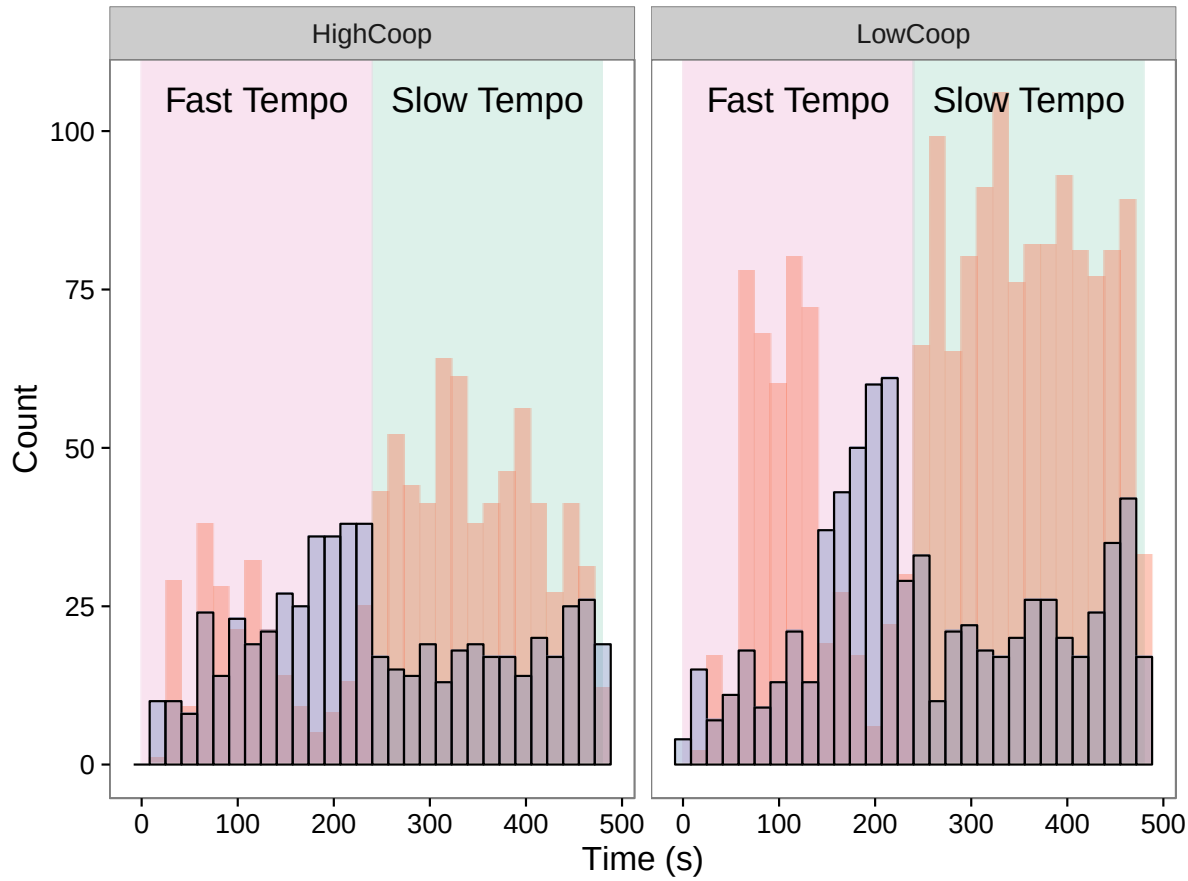

Figure 9. Number of participant requests (outlined bars) and agent requests (un-outlined bars) show that participants' requests peaked toward the end of their fast-tempo period.

The effects of participants' delayed requests also seemed to affect participants' and agents' acceptances. Contrary to Experiment 1, participants' mean acceptances were greater than agents' mean acceptances in Experiment 2 ($F(1, 34) = 15.97$, $p < 0.01$). It seems the low-cooperation agent was more badly in need of staff during participants' slow-tempo period in Experiment 2 compared to Experiment 1, due to participants' delayed resource requests and subsequent delayed use of staff at the start of their slow-tempo period, and thus agents' accepted less often in Experiment 2.

Without the benefit –or an example– of a requesting agent early in the trial, participants seemed focused on their scheduling rather than requesting staff from the agent.

This is especially true when comparing the timing of participants' requests in Experiment 2, which peak at the end of the fast-tempo period in Figure 9, and the timing of participants' requests in Experiment 1 which peaked earlier, halfway through the fast-tempo period (not pictured). Thus, participants' mean number of requests during an early fast-tempo may not have been sufficient. This would lead to requests increasing toward the end of the fast-tempo period to address buildup in the patient queue. Participants' tendency to adopt an autonomous strategy for assigning patients and staff to available rooms, meant they neglected the opportunity to gain staff to serve patients.

Despite this effect in Experiment 2, cooperation of the agent still had a positive effect on joint performance, supporting our initial hypothesis. The dyad in the high-cooperation condition had greater resource exchange between tempo periods compared to the dyad in the low-cooperation condition. This is reflected in the significant interaction term for cooperation level and tempo ($F(1, 34) = 4.46$, $p < 0.05$), the significant interaction term for scheduler, cooperation level, and tempo ($F(1, 34) = 17.02$, $p < 0.01$), and in Figure 8, where the differences between quadrants' left and right wide bars are greater in the left column ("HighCoop") compared to the right column ("LowCoop"). It is also noteworthy that compared to Experiment 1, and particularly in the high-cooperation condition, the fast-tempo period in Experiment 2 did not reduce participants' tendency to share staff, despite the agent demanding much less often.

It is tempting to compare performance in patients treated between the two experiments given their parallel design. However, plotting when agents released staff against when participants reached maximum queue length shows that in Experiment 1, even with earlier requests, participants were unable to treat the backlog of patients in their waiting room before the trial ended, mostly a function of the designed tempo patterns. Thus, the only conclusion we can make about joint performance in Experiment 2 is that it supported the

findings from Experiment 1 – cooperation affects joint performance. Joint performance was lower with the low-cooperation agent ($F(1, 34) = 12.24$, $p < 0.01$), and even lower when the participants initially interacted with the low-cooperation agent ($F(1, 34) = 5.63$, $p = 0.02$).

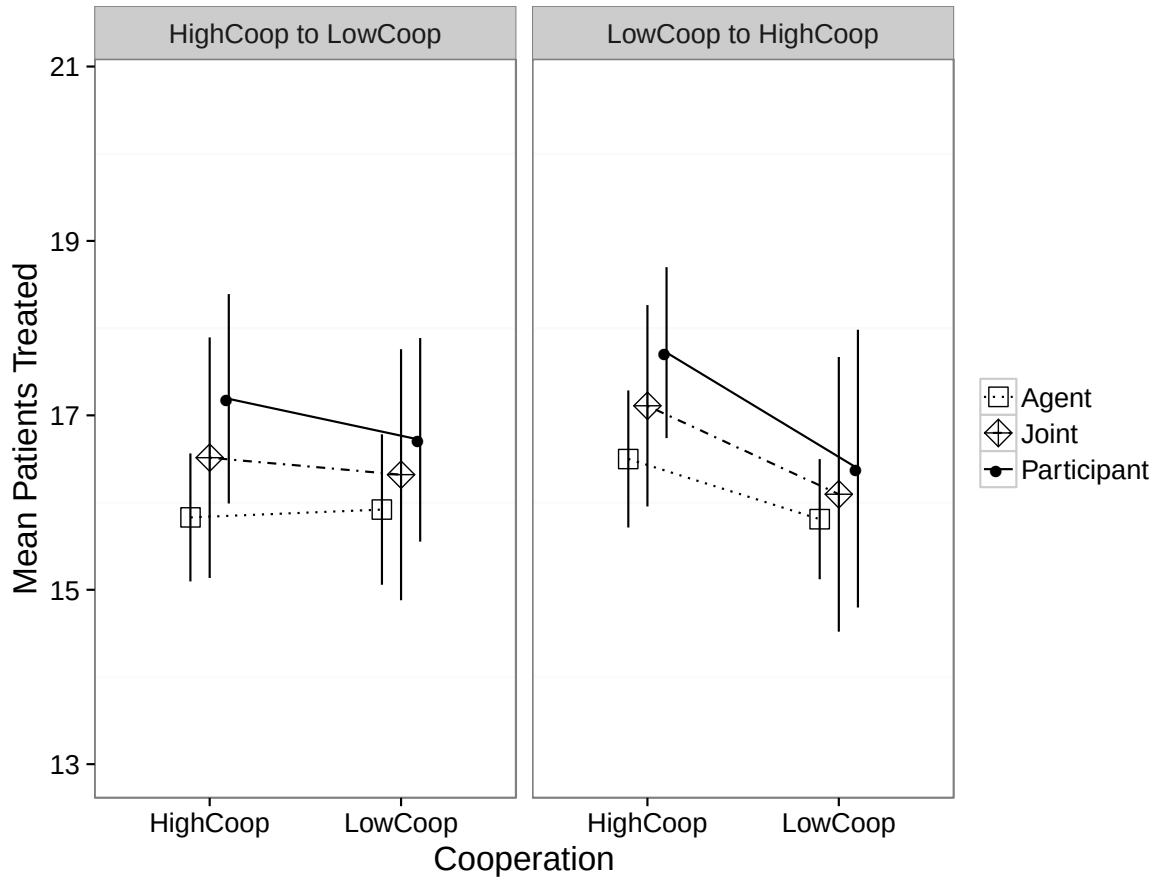

Figure 10. Mean patients treated with 95% CIs show individual performance and joint performance, with joint performance lower in the low-cooperation condition compared to the high cooperation condition.

To further investigate the interdependence of the agent and participant, a simple linear regression was calculated to predict joint performance based on participant performance. For Experiment 1 participants' performance accounted for a much smaller proportion of the variance in the joint performance ($F(1, 142) = 18.09$, $p < 0.01$, $R^2 = 0.11$) compared to Experiment 2 ($F(1, 142) = 273.00$, $p < 0.01$, $R^2 = 0.66$), supporting the idea that participants used a more autonomous strategy in Experiment 2.

3.5.3 Discretionary cooperation and rate of reciprocity

To compare overall discretionary cooperation between participants and agents, the number of acceptances made of "valid requests" were tallied for each participant. Valid requests were labeled by filtering out requests made when the other hospital did not have the resource being requested. The number of acceptances was then divided by the number of valid requests for each scheduler, and the resulting percentages were averaged across trials.

Comparing participants who experienced the fast-tempo first (Experiment 2, "Fast to Slow Tempo") and participants who experienced the slow-tempo first (Experiment 2, "Slow to Fast Tempo"), participant and agent dyads accepted more of each other's requests in the fast-to-slow tempo pattern ($F(1, 70) = 17.42$, $p < 0.01$) compared to the slow-to-fast tempo pattern (Figure 11). Agents accepted more of participants' requests than participants accepted agents' requests in both fast-to-slow and slow-to-fast tempo patterns ($F(1, 70) = 48.65$, $p < 0.01$), and the high-cooperation condition dyads accepted more requests than the low-cooperation condition dyads ($F(1, 70) = 141.762$, $p < 0.01$).

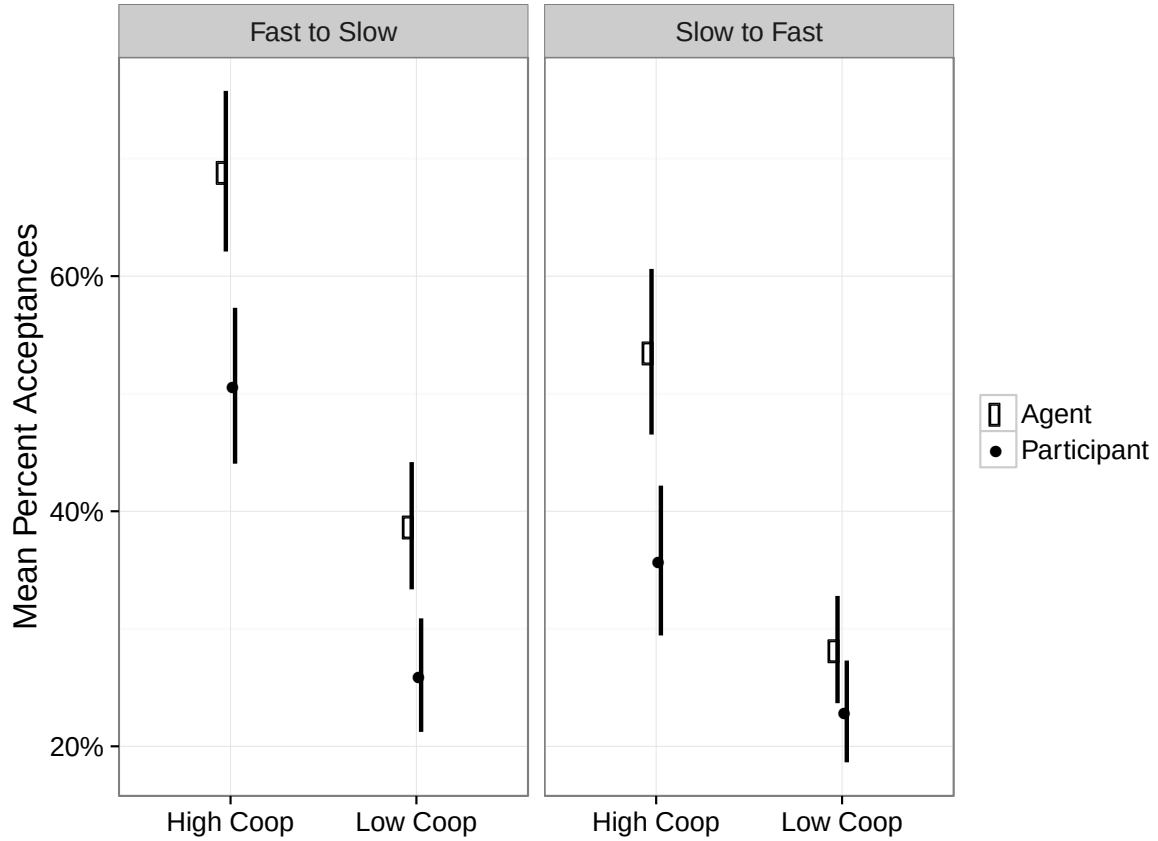

Figure 11. Comparison of mean percent acceptances of valid requests made shows that participants accepted less often than agents, and more acceptances were made in the high-cooperation agent condition than the low-cooperation agent condition.

Therefore, while the effects of agent cooperation were persistent across conditions, as were agents generally more cooperative than participants, the differences between the fast-to-slow tempo pattern (Experiment 2) and the slow-to-fast tempo pattern (Experiment 1) suggests they were different games rather than exact complements. That the agents accepted more of participants' requests than participants accepted of agents' requests opens the question, to what degree did reciprocity play in resource coordination?

Reciprocity was calculated as the percent difference beyond "matching" or returning the exact number of staff received from the agent (Cox, 2004) within the time frame of a complete experimental trial. The reason why returning based on prior exchanges, i.e., reciprocity through sequential turn-taking, is not used is because of the stochastic nature of the task at any given point of exchange – the available resources one could return depended

on the particular sequence of incoming patients, which was randomized between A or B type with 50 percent distribution. Thus, rather than calculating each sequence for each participant, for better generalizability, reciprocity is determined by calculating overall percentages at the end of each experimental trial.

Comparing reciprocity in Experiment 2 and Experiment 1 shows that in Experiment 2, participants demonstrated relatively high and relatively similar reciprocity (Figure 12, left column), with an average of 35.55% with the high-cooperation agent (*SD* = 49.75) and 37.81% with the low-cooperation agent (*SD* = 67.54). However, in Experiment 1, reciprocity differed between the two cooperation conditions (Figure 12, right column). Participants in the slow-to-fast condition returned fewer staff to the high-cooperation agent than what they received (*M* = -18.69%, *SD* = 20.66%); where 0% reciprocity means they matched the agent's number of staff shared. One explanation for this difference is because high-cooperation agents were more generous, so there was a steeper baseline for matching the number of resources given by the high-cooperation agent.

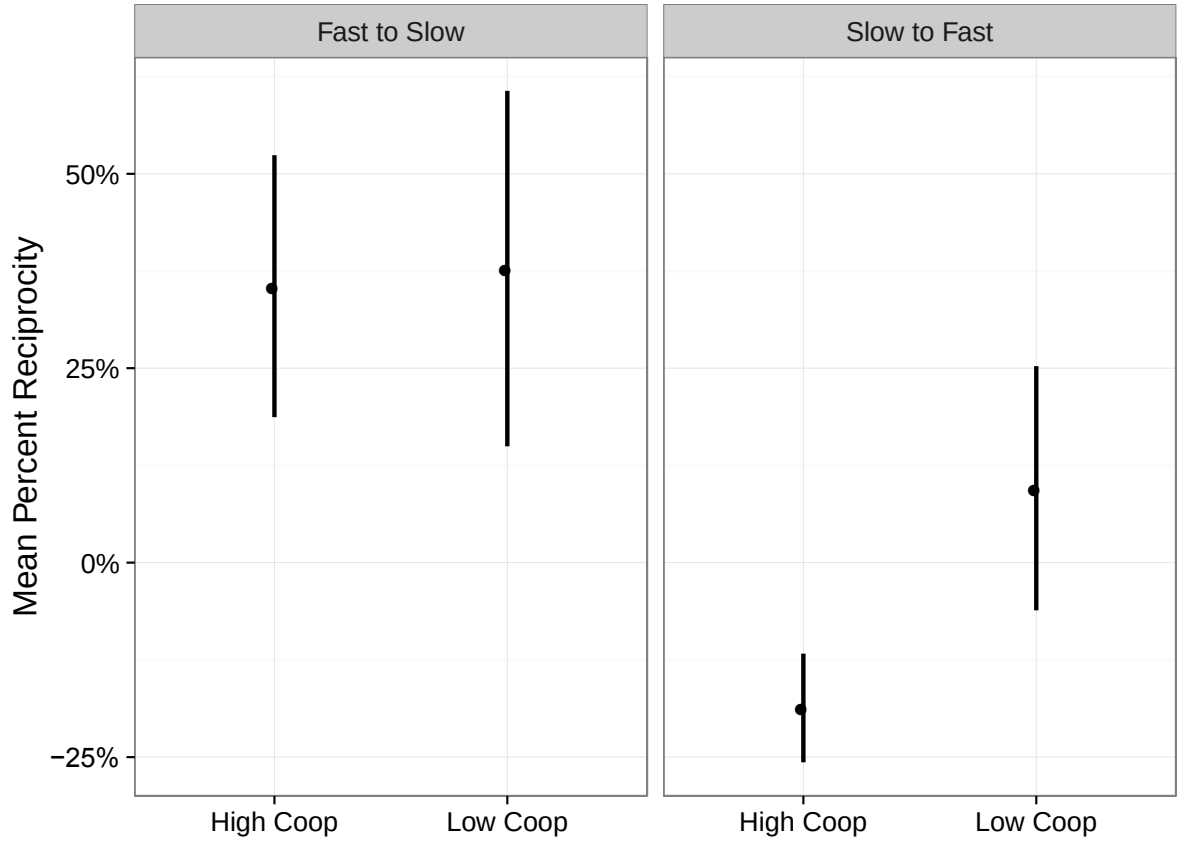

Figure 12. Mean reciprocity in Experiment 2 (left) and Experiment 1 (right) show that reciprocity with the two cooperation agents was relatively similar in the fast-to-slow tempo, whereas in the slow-to-fast tempo people did not reciprocate with the high-cooperation agent, but did reciprocate with the low-cooperation agent.

Therefore, with the low-cooperation agent they reciprocated more easily, returning a number of resources greater what was received (M = 9.57, SD = 46.4), although still lower than participants in the fast-to-slow condition. These observations are supported by the significant difference found between the two tempo pattern conditions ($F(1, 68) = 17.39$, $p < 001$), between cooperation levels ($F(1, 68) = 6.18$, $p = 0.02$, and their interaction term ($F(1, 68) = 4.48$, $p = 0.04$).

The lack of a large difference between cooperation conditions in the fast-to-slow tempo condition indicates reciprocity had less of an effect on resource exchange behaviors between schedulers. One explanation for this is that in an interaction structure where resource exchange is negotiated, acceptances can only come from requests initiated. Participants in the

fast-to-slow condition had more time to request staff following an initial fast-tempo period, compared to the slow-to-fast condition participants who had less time to request staff following their fast-tempo period in the second half of the trial. In contrast, the large difference between cooperation conditions in the slow-to-fast tempo reflects that people were able to use the high-cooperation agent's staff effectively, leading to higher joint scores than with the low-cooperation agent. Returning a greater number of staff to the low-cooperation agent than what they received may have led to the group's lower joint scores – perhaps both human scheduler and low-cooperation agent would have benefitted more from participants withholding slightly more of the staff obtained from the low-cooperation agent in the slow-to-fast tempo condition.

3.6 General Discussion

Both experiments confirm that cooperation with automation is an important complement to reliance on automation. People and automated agents were placed in parallel roles to focus on the need to coordinate in a dynamic task environment through cooperation. While sharing resources potentially produces conflicting goals, in this study, demand was engineered to minimize conflict. From the researcher's perspective, there was a relatively clear coordination solution – schedulers in the slow-tempo would be able to share staff with their counterparts experiencing a fast-tempo period. This allowed us to test how participants responded to agents expressing different levels of cooperation, and to motivate future studies on the social or affective influences of cooperation, rather than only the information processing components of the task. Overall, both experiments showed agent cooperation affected participant cooperation, supporting our general hypothesis that agent cooperation is an important construct in human-agent coordination, beyond automation reliability and its effects on reliance and compliance.

Although the goal was to "treat as many patients as possible," participants could take a defensive approach (Stephens et al., 2011), requesting staff without returning them, to maximize patients treated in their own hospital. However, such defensive behavior would undermine cooperation and reciprocity (Axelrod & Hamilton, 1981), and in the microworld, lead to fewer patients treated. Though resource-sharing decisions may have been governed by the instrumental value of staff resources (Molm et al., 2007), the environment demand was not made explicit to participants. Therefore, the act of sharing resources demonstrates their trust in the agent. Accepting requests came at the cost of losing spare staff, reducing their margin of maneuver, and risking that the agent would not reciprocate and return staff when needed. In both experiments, less cooperative agents induced less cooperative behavior in the participants and more cooperative agents induced more cooperative behavior in the participants.

The hypothesis for Experiment 1 on cooperative reciprocity was thus supported. Although the rate of reciprocity was inconclusive in Experiment 1 due to the structure of the negotiated interaction, the cooperative nature of the two agents affected both dimensions of participants' cooperation – resource-sharing and resource-requesting behaviors. Participants engaged in more productive exchanges – successful requests – with the high-cooperation agent than with the low-cooperation agent, despite the low-cooperation agent requesting staff from participants more often. This suggests people returned similar cooperative behaviors of the agents. In addition, participants who previously experienced a low-cooperation agent and were working with the high-cooperation agent demonstrated even more productive exchanges than participants who experienced the high-cooperation agent first. It may be that participants were better at timing their requests in later trials to coincide with when the agent had available staff, but there was no evidence of learning effects. It may be that the high-cooperation agent surpassed low expectations formed with the low-cooperation agent,

indicating the interaction history influenced decisions to cooperate. This latter explanation and the higher joint score with the high-cooperation agent support the idea that to maintain resilience, systems must continually invest in a renewal of shared goals and a willingness to accommodate (Woods, 2004).

The hypothesis for Experiment 2 – that if people's initial exposure to an agent is in a highly demanding situation, then the effects of agent cooperation level will be more prominent – was partially supported by participants' mean acceptances, particularly in the fast-tempo, high-cooperation condition. However, the reciprocity rates show that participants withheld more staff given to them by the high-cooperation agent, which led to higher joint scores compared to participants coordinating with the low-cooperation agent. This suggests the hospital microworld mechanics may not have supported a true one-to-one resource coordination scenario, and that while reciprocal cooperation may have been present in resource exchanging behavior, it was not necessarily the case that reciprocity as measured by a positive amount returned compared to what was provided would have led to higher joint scores.

As alluded to earlier, the initial fast-tempo period in Experiment 2 as well as the need to make requests to receive staff, may have affected both participants' rates of reciprocity and their ability to coordinate. Overwhelmed participants may have taken on more autonomous rather than cooperative requesting strategies (not requesting and requesting, respectively). Rather than anticipating demand and making sufficient early requests, participants' delay led to a backlog that undermined their ability to accept requests as agents transitioned into fast-tempo (especially the low-cooperation agent). However, when prompted, people still responded in a cooperative way that contrasted their relatively autonomous requesting strategy. This suggests cooperation may have two dimensions – proactive and responsive actions. Given that participants started with a medium-tempo margin in each trial, and were

free to request staff, it remains a question why participants did not anticipate demand in subsequent trials. Certainly expertise with such demanding situations in actual work environments might ameliorate the tendency towards an autonomous, reactive (Hollnagel, 2012) response. However, these results are consistent with the general error tendency of cognitive tunneling when people confront unexpected high-demand situations (Woods et al., 1994). In summary, this autonomous behavior was limited to reducing participants' requests for staff, and did not reduce their tendency to accept requests.

Previous work in human-automation cooperation has focused on performance, with adjustable autonomy typically referring to better coordination (Zieba et al., 2010) or better management of functional dependencies. That speed of assigning staff was an important factor for this particular task might make this study a strange example of a joint human-agent task; automation is known to be much better at this than people (Fitts et al., 1951). However, it is not always the case that automation should be used, when the costs of a disengaged human counterpart would be greater than the efficiencies of implemented automation (Kirlik, 1993). Our study focuses on the social processes of coordination in resilience, in circumstances where reciprocity and goal tradeoffs feature more prominently than reliability and mode management. Doing so departs from the idea that teams should be formed purely based on complementary abilities, rather than how those team members interact. It also challenges the tempting idea that if functions can be automated then they should be automated.

Establishing trust between people and autonomous agents may be one of the most daunting problems for the success of human-automation teams (Groom & Nass, 2007), which is important for cooperation and maintaining margins of maneuver. Working toward this ideal, we investigated a shared-resource activity between people and automation, where social processes could potentially limit joint performance. As Groom and Nass (2007)

observe, future research intent on developing human-robot teams must go beyond technical performance to address the social and organizational qualities that make a successful teammate.

3.7 Limitations and Future Directions

Microworlds have the advantage of mimicking interactive situations observed in field studies but with a greater degree of experimental control (Gonzalez et al., 2005). This experimental control is particularly valuable when investigating the dynamics of interdependent behavior. Microworlds make it possible to manipulate the behavior and interdependencies of dyads, the simplest unit of agent networks in complex sociotechnical systems, but such interdependencies can also lead to challenges in the analysis.

In this study, the high-cooperation agent's requesting behavior could be affected by participants who were slower or chose not to use their staff because the high-cooperation agent was designed to check if participants' staff were unused prior to requesting them. The low-cooperation agent's requesting behavior could be affected by how quickly it ran out of staff it needed, which could be influenced by staff it did or did not have due to the participant's actions. Both high-cooperation and low-cooperation agents' acceptances depended in part on when participants made requests. This led to challenges in interpreting the results. However, this interdependence approach echoes the challenge of designing strategies to combat conflict in such joint tasks, where a person's decisions depend on the agent's decisions, which depend on the person's decisions, and so on (Schelling, 1960). Most studies seek to avoid interdependencies, but a systems perspective that considers these interdependencies is becoming more important with the changing role of automation.

In practice, it may be difficult to draw conclusions about the impact of cooperative behavior on performance outcomes; it is possible that as cooperative activity goes up, productivity decreases or remains unchanged, or other important measures are affected, such

as worker turnover and quality of work life. Cooperation may create value where there is none, and the quality of the output is not necessarily measured in terms of responding to demands more quickly (Wicks, Berman, & Jones, 1999), as represented in this microworld. In other words, increased cooperation might not increase performance in routine situations, even if it enhances resilience in high-demand situations. As a consequence, organizations might neglect the importance of cooperative automation.

It is challenging, but critical, to identify performance measures that emphasize priorities consistent with cooperative control. In practice these priorities can manifest as micro-decisions to cooperate that extend to macro outcomes for a system of interconnected agents. In designing cooperation for resilience, the next study will avoid reactive responses by exploring a different social exchange structure, moving from negotiated exchange, where exchange decisions are made jointly, to reciprocal exchange where decisions are more unilateral, like in altruism (Molm et al., 2000). Such a design could refocus human-agent interactions from interrupting to request staff, a structure with more reactive affordances, to interrupting with needed staff, a structure with more proactive affordances. A key question would be to understand how to foster proactive behavior that helps systems maintain margins of maneuver, and to understand what causes people to ignore or avoid early opportunities that could enhance cooperation.

3.8 Conclusion

Cooperation is central to resilience because it enables networks of interdependent agents to pool staff and accommodate a greater range of surprises. People can become less cooperative when interacting with less trusting and inconsiderate agents, and more cooperative when interacting with more trusting and considerate agents, and cooperative strategies might be expressed in both proactive and responsive behaviors. These behaviors are qualitatively different than those observed in investigations of supervisory control

automation. This study's approach thus departs from the typical supervisory control approach to human-automation interaction, and acts as a starting point for future studies to explore how cooperation in human-agent interactions can enhance system resilience.

Chapter 4 Part 2: Effects of Reciprocal Exchange on Human-Agent Cooperation

4.1 Lessons Learned from Part 1

Findings from Part 1 (Chapter 3) show that people's cooperative behaviors were differently affected by agent cooperation and by workload. Although people were generally more cooperative with a high-cooperation agent, and less cooperative with a low-cooperation agent, requesting behaviors were delayed during the high-workload situation while their accepting behaviors were not as affected. This suggests cooperation may have a dimension of proactive (requesting, in this case) to responsive exchange behavior. People's proactive resource-acquisition performance suffered during the high-workload condition because their requests were delayed. This is consistent with the general error tendency of cognitive tunneling when people confront unexpected high demands (Woods et al., 1994). It thus seems that high-workload led participants to narrow their focus on scheduling in their own hospital, failing to request staff, though they were still willing to provide staff when prompted.

High-workload is often met by suggestions to reduce workload, such as increasing staff or implementing automation. However, high-workload situations are often unexpected and unavoidable in complex systems, where activity ebbs and flows, with periods of slower self-paced activity interspersed with higher-tempo, externally-paced demands (Rochlin et al., 1987). Automation is often used to shift workload or tasks from people to machines, but a critical feature of well-integrated work is not simply a reduction of workload by eliminating tasks. Instead, well-integrated work considers how automation impacts low-workload and high-workload periods, and how automation promotes cooperation and people's ability to manage workload, by increasing operating margins given resource limitations (Rankin, Lundberg, Woltjer, Rollenhagen, & Hollnagel, 2013; Woods et al., 1994). People often need to fill the gaps that system design or system failures produce, and keeping them in the loop with equal authority better enables them to do so.

Rather than simply reducing workload by increasing automation, an alternative approach could consider changing the social exchange structure, from one that requires action-response in an exchange, to one that focuses on resource provision in exchange. Doing so removes potential complacency in the resource-sharing task – to over-rely on automation to prompt of its needs and focus on self-needs – and increases accountability through initiative in resource-sharing. Such a shift in focus, based in an affordance of the social exchange structure (Kelley et al., 2003), is a novel intervention for human-automation interaction design that may improve cooperative resource-sharing to enhance resilience, rather than encourage defensive or autonomous resource management.

4.2 From Negotiated Exchange to Reciprocal Exchange

In Part 1, a negotiated exchange structure was tested, with a proactive resource acquisition (request resource) and responsive resource provision (accept or deny request) design. A different negotiated exchanged study could have tested a proactive resource provision (offer resource) and a responsive resource acquisition (accept offer design). A reciprocal structure, on the other hand, involves input from one member of the exchange, the initiator (Molm et al., 2007). Thus, the other member has no control in the decision, as would be the case if resource-sharing were fully automated (only the agent determines resource allocation) or fully manual (only the person determines resource allocation). However, because both members equally lack control in acquiring staff, the partnership is still an interactive control relationship, rather than a supervisory control relationship. In the negotiated exchange, actions are thus either “proactive” or “responsive” because both members participate in the decision and resulting outcome, whereas in reciprocal exchange, actions are either “proactive” or “passive”, because only one member – the initiator – can take action toward the immediate outcome of an exchange decision.

Previous research in social exchange theory suggests that cooperation may improve in a reciprocal exchange structure compared to a negotiated exchange structure because reciprocal exchange produces stronger trust and commitment between partners than in negotiated exchange. The reasoning behind this is that the symbolic value of pro-social behavior, an action that signal trustworthiness, increases when a decision is made without expected return; it places the initiator in a more vulnerable position (Berg et al., 1995; Molm et al., 2000). Due to the relationship between reciprocity and cooperation established in prior work, it thus seems likely that reciprocity will have a greater effect in a reciprocal exchange structure, with people reciprocating more than in negotiated exchange structures. Furthermore, in high stress environments, people's subconscious processing of risk and subsequent engagement in either defensive or social engagement strategies may be affected by exposing participants to an environment of positive, prosocial behavior (Porges, 2001, 2007). This supports the work in social exchange theory that suggests a proactive, giving agent may induce people's social engagement and prosocial resource-exchange behaviors even in a risky and high-stress environment.

People's trusting dispositions, or propensity to trust, has been shown to influence their trust-related decisions (Hancock et al., 2011; Kim, Ferrin, & Rao, 2008; Parks et al., 1996; Robert, Denis, Hung, Dennis, & Hung, 2009). It is thus expected that people with a higher propensity to trust will be more cooperative in a reciprocal exchange structure. Since individual differences are not the focus of this dissertation, measuring people's propensity to trust was used as a control to account for potential outliers or an unexpected trend in the data. In the following study, propensity to trust was measured using 20 items from the Propensity to Trust Survey (Evans & Revelle, 2008), which is shown to be effective at measuring trust in economic situations compared to other established scales for dispositional trust (Rotter, 1967,

1971). For more details on how these questions were adapted for this dissertation, see Appendix A: Propensity to Trust Questionnaire.

In addition to the affective potential of reciprocal exchange, its structure focuses attention on the other's needs rather than to self-needs. Instead of proactive attention to their self- needs (requesting staff) and responsive attention to agent needs (accepting/denying requests), participants would need to be proactive about the agent's needs (giving staff) and passive about their own needs (no control). In other words, initiating interactions changes from "what do I need?" to "what can I provide?" As such, high workload and subsequent cognitive tunneling in the hospital-scheduling task may be reduced by changing the interaction structure to reciprocal exchange, shifting the focus to the resource-exchange task, and reducing the additional workload imposed by negotiated exchange. Such a structure may be effective particularly in situations where one entity's peak workload corresponds to another's trough, with the trough allowing spare capacity to look out for the needs of the other.

Along those lines, people's perception of task interdependence may also impact cooperative behavior (Staples & Webster, 2008). Perceived level of task interdependence has been found to affect decisions to cooperate (Martin, Gonzalez, Juvina, & Lebiere, 2013). Furthermore, when one party's actions are understood to be contingent on another's behavior, perceived interdependence thus reduces risk and encourages cooperation (Cropanzano & Mitchell, 2005). In the following study, task interdependence was measured through five items adapted from Staples and Webster's (2008, p. 640) six-item scale for task interdependence, based on work by Bishop and Scott (2000), and Janssen (1999) (Appendix B: Task Interdependence Questionnaire).

This social exchange approach in Part 2 continues to explore an aspect of human-automation interaction that departs from more traditional approaches of function allocation in

supervisory control. A supervisory control approach might burden an automated agent to both push and pull relevant task information, and burden people with remaining vigilant of others' needs in the face of cognitive tunneling. Rather than automating everything that can be automated, which can lead to operator complacency (Kirlik, 1993), the following study examines a new interaction structure that would be less relevant in more defined supervisory control relationships, but may provide useful insights for designing interactive control relationships in dynamic, coordinative tasks that support resilience.

4.3 Research Questions

In the context of this research, changing the joint work from a negotiated exchange to a reciprocal exchange may improve participants' cooperation due to the affordances of reciprocal exchange that enhance proaction. In manipulating the exchange structure, the expected outcome is increased resource exchange in Part 2 (comparing gifts in Part 2 to requests accepted in Part 1), and subsequently improved ability of the human-agent system to adapt quickly to unexpected demand. Two versions of the automated agent were also tested in Part 2, a low-cooperation and high-cooperation agent. It was expected that the differences in cooperation between participants who interact with the high-cooperation agent and participants who interact with the low-cooperation agent will be greater, due to the increased symbolic value of unrequested giving.

4.4 Overview of Microworld Environment for Reciprocal Exchange

The technical equipment and interface design were different in Part 2 from Part 1, due to the differing nature of the task, as well as factors unrelated to experimental design, including an improved interface resulting from design iteration and technical considerations for scaling the microworld for future studies. The microworld for reciprocal exchange was developed in Java and XML with the Java Development Kit 7. A 15" laptop computer with Java Runtime Environment 7 and a standard computer mouse were used. As in Part 1,

participants acted as hospital schedulers, whose task involved assigning patients and staff to hospital rooms and coordinating shared staff resources with a neighboring hospital, managed by an automated agent. Interactions with the agent were unilateral and limited to giving staff resources and accepting staff resources.

4.5 Experimental design

The study tested different levels of agent cooperation on participants' giving behavior and joint performance as a between-subjects variable; 25 participants experienced a high-cooperation agent, and 25 participants experienced a low-cooperation agent. The fast-to-slow tempo pattern was tested, in which participants experienced a fast-tempo period followed by a slow-tempo period, the same as in Experiment 2, Part 1. Therefore, the within-subjects variables in this experiment included trial (trial 1 or trial 2), tempo (fast-tempo period or slow-tempo period) and scheduler where applicable (participant or agent).

4.5.1 Study participants

A total of 50 participants were recruited from a midwestern university community, through online job and volunteer postings and campus fliers near the study site. Criteria for participation included normal or corrected-to-normal vision, between 18-65 years old, access to an email address, and able to use a standard computer keyboard and mouse. Interested participants who contacted the researcher were asked to confirm they met the criteria and provide their availability. Participants were then randomly assigned to condition groups for the experiment. Following the experiment, participants were compensated $10 for completing the study.

Participants' self-reported genders were 42% male, 56% female, and 2% other; ages ranged from 18 to 37, with a median age of 22 and a mean age of 23. Most participants reported using the computer daily (96%), and 38% reported using the computer for playing video games. Participants' mean scores on questionnaire items relating to their trust

dispositions ranged from 3.05 to 4.68, with an overall mean of 4.08, out of a scale from 1 to 6, with 6 meaning higher trust (Evans & Revelle, 2008). Participants' understanding of task interdependence was also measured, and they generally understood that their task was interdependent with the agent's task. On a scale of 1 to 6, with 6 meaning highly interdependent on the task interdependence questionnaire items (Staples & Webster, 2008), participants' median score was 4.6. For a full list of the questionnaire items and demographics collected, see Appendix A: Propensity to Trust Questionnaire, Appendix B: Task Interdependence Questionnaire, and Appendix C: Demographic Questionnaire.

4.5.2 Independent variables

Agent cooperation, operationalized in the agents' staff-sharing rates and behaviors, was tested as a between-subjects variable. Each participant played two trials, so "trial" was included as a within-subjects variable. As in Part 1, Experiment 2, within each trial there was one fast-tempo period, and one slow-tempo period, so "tempo" was tested as a within-subject variable. "Scheduler" was a within-subject variable, to distinguish between the participant and the automated agent. Therefore, a mixed between and within 2 x 2 x 2 x 2 designs, and 2 x 2 x 2 designs, tested agent cooperation level (high or low cooperation) as a between-subjects variable, with within-subjects variables including trial (trial 1 or trial 2), tempo (fast-tempo period or slow-tempo period) and scheduler where applicable (participant or agent).

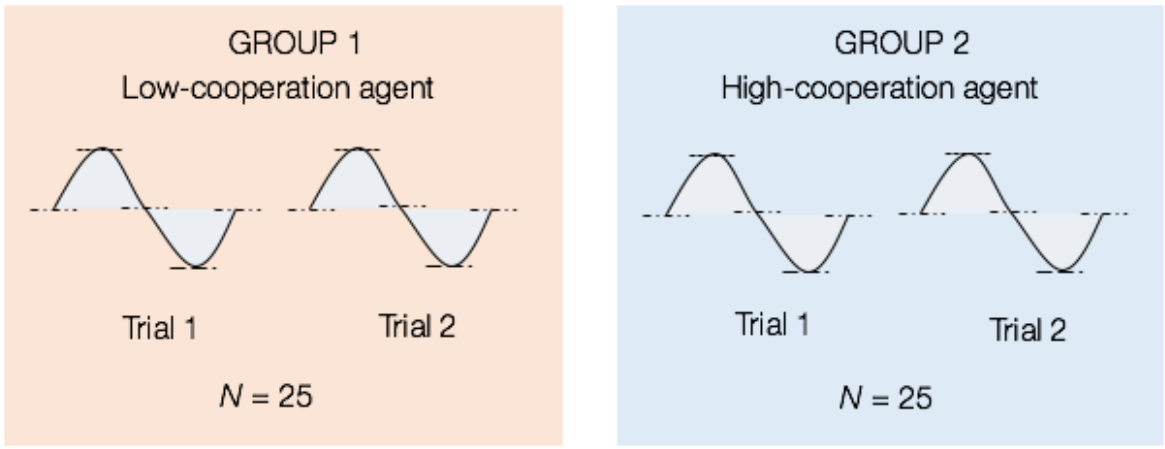

Figure 13. Two groups of 25 participants: one group experienced the high-cooperation agent, the other group the low-cooperation agent, both a fast-to-slow tempo sequence, while their agents experienced the slow-to-fast tempo sequence.

4.5.3 Tempo pattern

As in Part 1, Part 2 participants and agents experienced complementary tempo patterns. However, participants only experienced the fast-to-slow tempo (for more details see section 3.3.3). Testing this tempo pattern focuses on the staff-sharing response of participants in the second half of the trial, following experience with a staff-sharing agent in the first half of the trial. In cooperative exchange, the first move or first impression is an important factor of the subsequent relationship (Axelrod, 1984). Thus, a slow-to-fast tempo pattern may be more susceptible to exposure order effects from starting in an environment where people need to proactively provide resources. For a note on why a baseline tempo pattern was not tested, see (Appendix D: Baseline Tempo).

4.5.4 Agent cooperation levels – high-cooperation and low-cooperation

Similar to the high- and low- cooperation agents in Part 1, Part 2 tested two cooperation agents, also named high- and low-cooperation. Agent cooperation was operationalized as high-cooperation or low-cooperation depending on the degree to which its behaviors signaled joint outcome or individual outcome, respectively. Although the behaviors of the two agents in Part 1 and Part 2 seem similar, due to the changed structure in Part 2,

agent behavior was different in a few critical ways. The Part 2 agents' behaviors are summarized in Table 3 below.

Table 3. Agent resource-sharing and giving behavior by cooperation level

Agent Cooperation Level	Giving Rate	Giving Behavior
High cooperation	100 % with 1-2 patients 75 % with 3-4 patients 50 % with 5-6 patients	1. If participant spare staff is at 0, 2. If agent has that unassigned staff, 3. Give that staff at pre-specified giving rate
Low cooperation	50 % with 1-2 patients 25 % with 3-4 patients 0 % with 5-6 patients	1. If participant spare staff is at 0, 2. If agent has any unassigned staff, 3. Give randomly at pre-specified giving rate

Note. Agents' giving rates were keyed to the number of patients in the agent's waiting room, and to a color status in the waiting room: green for 1-2 patients, yellow for 3-4 patients, and red for 5-6 patients.

Due to the structure and constraints of the microworld designed, cooperation entailed providing useful resources at the appropriate time. However, agents still needed to treat patients in their own hospitals to contribute to a higher joint score. Therefore, agent cooperative behavior was necessarily tied to the status of their own hospital.

To make the high-cooperation agent more attuned to joint performance, and more cooperative, the high-cooperation agent was designed to be more discriminate and considerate; it would check if the participant had any staff at zero and provide that specific staff if it had it. Then, the high-cooperation agent provided participants staff at a rate of 100% when its own queue was green, 75% when its queue was yellow, 50% when its queue was red. While this rate was higher than the low-cooperation agent's rates, its discriminate giving meant that it had fewer chances to provide a specific staff resource participants needed, and thus the high-cooperation agent provided approximately the same number of staff resources to participants as the low-cooperation agent.

The low-cooperation agent was positioned to be more attune to its individual performance, in that its giving behavior was less risky for itself, emphasizing individual outcome. The low-cooperation agent's giving rate was less than the high-cooperation agent's,

and it gave more indiscriminately, without considering participants' needs as carefully as the high-cooperation agent. While the low-cooperation agent also checked to see if participants needed staff, i.e., any staff values at zero, it gave random staff not currently in use 50% of the time when their queue was green, 25% when yellow, and 0% when red. This added check of participants' staff values in the low-cooperation agent was included following pilot studies that found a low-cooperation agent without this check tended to overwhelm participant partners with staff resources.

To prevent agents from excessively returning staff to participants, if agents received a resource from participants, it first checked to see if it could use the resource before enacting its giving behavior decision cycle. Both agents also had time delays to simulate the pace of a human player, established during pilot testing. Before running through an entire decision cycle, agents would delay 4 seconds, with an added 1-second delay when collecting staff from a room. This delay avoided continuous interruption of the participant and allowed a window between assignments for agents to have excess staff that would ensure interaction.

4.5.5 Dependent variables

Because we wanted to know how agent cooperation would influence a person's cooperation in a reciprocal exchange structure, the dependent variables in this study measured different parts of cooperation: cooperation behavior, cooperation process, and cooperation outcome. To measure cooperation behavior, we considered the number of staff participants transferred to the neighboring hospital agent. To capture the cooperation process, we considered both the temporal pattern of staff transfers as well as the number of staff participants reciprocated to the agent. Both cooperation process measures were derived from the cooperation behavior measure. Finally cooperation outcome was measured, which was the number of patients treated in both hospitals. This approach to understanding cooperation considers how micro behaviors develop into patterns and ultimately influence outcome. Table

4 summarizes these parts and their corresponding measures. Because of the interdependence of the joint task, agents' resource-sharing behaviors and process were also measured to help contextualize participants' experience.

Table 4. Dependent Variables

	Measure	Unit of Analysis
Cooperation Behavior		
Providing staff to neighboring hospital	Number of Staff	Participant
Cooperation Process		
Providing staff based on demand pattern	Temporal pattern	Participant
Reciprocated staff provided by agent	Relative difference of staff provided	Participant + Agent
Cooperation Outcome		
Joint performance	Number of patients treated	Participant + Agent

4.6 Procedure

Participants' consent was obtained first, then they completed a propensity to trust questionnaire, were trained on the microworld task, and completed the experimental trials. After the experimental trials, they completed a task interdependence questionnaire and demographics questionnaire. In Part 2, only two experimental trials were conducted instead of four; Part 1 demonstrated two were sufficient exposure to capture participant and agent behaviors, and the within-participant variable "cooperation order" was not tested in Part 2.

Table 5 describes the main actions of the scheduling task from the participant's point of view. Level 1 refers to the right hand control panel options in Figure 14.

Table 5. Participants' main actions and navigation pathways in reciprocal exchange

Level 1	Level 2	Level 3	Level 4
Actions and Pathways Related to Scheduling			
Assign: Patient A/Patient B	*(Select available room 1-6)*	--	--
Assign: Doctor/Nurse/Surgeon	*(Select available room 1-6)*	--	--
Actions and Pathways Related to Resource Exchange			
Give Resource	*Give: Doctor Nurse Surgeon*	"Are you sure you want to send a (Doctor/Nurse/Surgeon) to the neighboring hospital?"	*Yes*
			No
"The neighboring hospital gives you a (doctor/nurse/surgeon)."	*OK*		

Note. Level 1 indicates the first level of options and the subsequent levels describe pathways for the options selected. Italicized texts refer to button options and quotations refer to pop-up window communications. Two dashes indicate levels not applicable to that pathway.

Interface highlighting aided participants in the microworld (Figure 14). For example, unavailable options such as patients, staff, or rooms occupied were greyed out, and the system did not allow erroneous assignments. Once a patient and staff were assigned to a room, "patient treatment" began automatically, lasting 60 seconds. After treatment, participants needed to click a "Collect Resources" button that appeared on top of the respective room to free the room and staff for reassignment (action not included in Table 5).

To give staff, participants could select "Give Resource" from the bottom of the right hand control panel, then select nurse, doctor, or surgeon. After confirming the decision (Table 5, Level 3) the resource would transfer to the agent's hospital and the staff numbers in both hospitals would update. Because reciprocal exchange is unilateral decision-making, agents and participants alike had no choice but to receive the staff given to them. However, interface design ensured participants were aware of these interactions. When the agent gave the participant a resource, an immediate interruption (McFarlane & Latorella, 2002) via pop-

up window appeared with the message, "The neighboring hospital sends you a (nurse/doctor/surgeon)" and participants needed to click “OK” button on the pop-up window before continuing (Table 5, Level 1 and 2). This ensured that participants knew when staff transferred from the neighboring agent’s hospital to their hospital. In Part 1 where participants requested staff, participants first selected the option to request, then indicated which resource they were requesting. In Part 2, participants selected the option to give, and then indicated which resource to give.

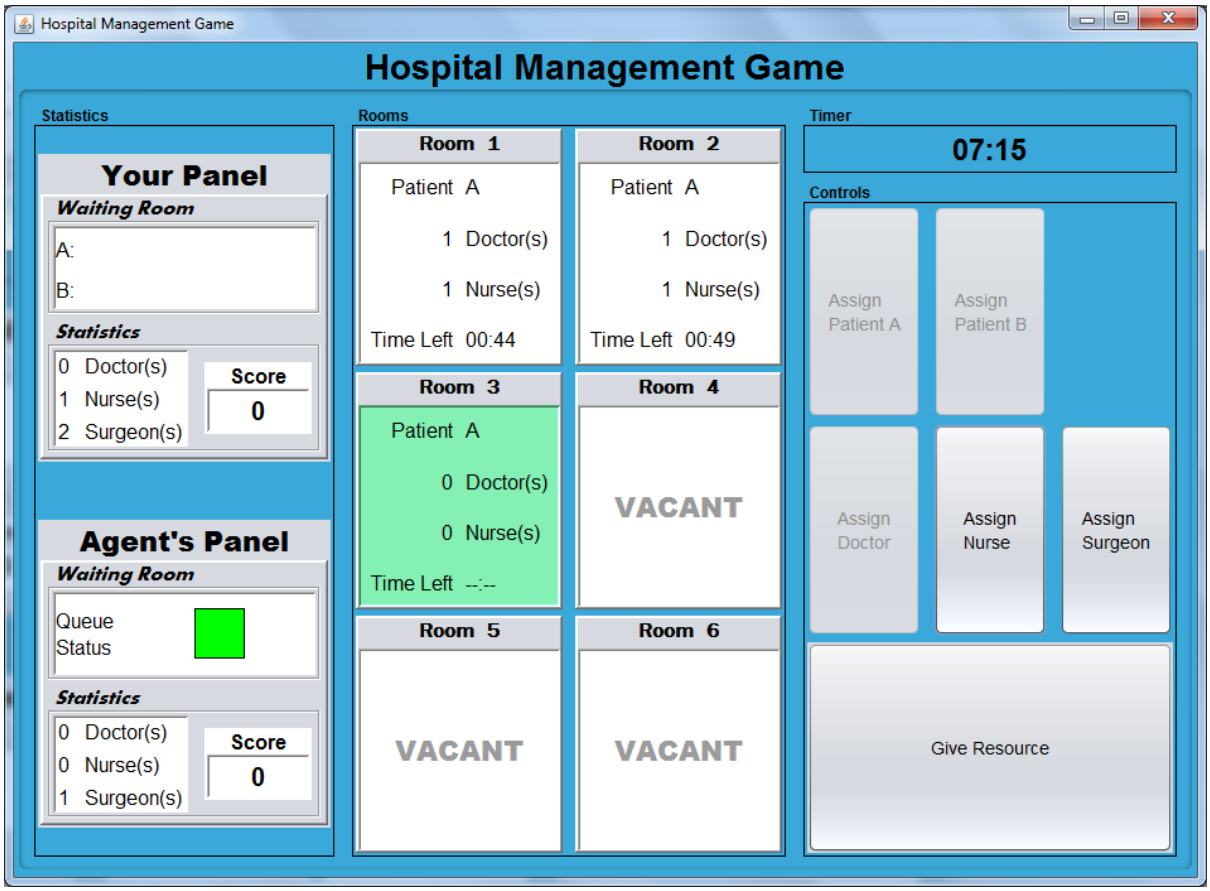

Figure 14. Screenshot of the microworld design used in Part 2 shows the participant in the middle of assigning a resource to Room 3, which is already assigned a Patient A and is highlighted to demonstrate which room can take the selected resource.

The automated agents only assigned patients to rooms when all staff were available. This prevented the agents from accidentally hoarding staff by assigning a patient and one staff resource to a room that could not begin treatment until the second staff resource is

available. If there were no patients to assign with the accompanying full set of appropriate staff, then both agents would enter the giving behavior decision cycle (Table 3).

4.7 Data Labeling and Analysis

Prior to analysis, two participants' data were removed; for more details see Appendix E: Two Outliers. The analyses conducted considered cooperation in terms of resource-sharing actions, in the context of the shared goal and task environment with shifting priorities over a specified period of time. ANOVA was used to compare the joint scores and number of staff exchanged between groups (between-subjects), number of staff given in the slow- and fast-tempo periods (within-subjects), and in each trial (within-subjects).

An ANOVA was also applied to the dependent measure of reciprocity, which was calculated as a percent difference from the overall number of staff returned to the agent per trial. As discussed in section 2.4.2, values above 0% were therefore considered the degree to which reciprocity was demonstrated (Cox, 2004); values below 0% were not considered to be reciprocity, and 0% meant resources lent were returned. For clarity, the criteria for reciprocity are measured as a separate construct from cooperation, but due to the relationship between reciprocity and cooperation established in prior work, they are considered in conjunction with cooperative behaviors to assess its role in explaining cooperation and joint performance.

Finally, an independent two-group t-test was used to compare joint scores between Part 1 and Part 2 due to the differing sample sizes (N = 18 per group and N = 25 per group, respectively). Analyses were conducted in R and using the 'stats' package (R Core Team, 2014); data frame manipulations and descriptive statistics were conducted using the 'dplyr' package (Wickham & Francois, 2015); figures were created using 'ggplot2' (Wickham, 2009).

4.8 Results and Discussion

In reporting results, all significant terms and insignificant main effects are included; insignificant interaction terms are not included. Propensity to trust and perception of task interdependence were included as covariates for each analysis reported below and were not found to account for significant portions of the variance, so their results are included as part of the sample population description (Section 4.5.1). All figures are labeled from the participants' perspective.

4.8.1 Joint performance

Mean joint scores were not significantly different between cooperation conditions ($F(1, 48) = 0.06$, $p = 0.81$). Higher scores were reached in trial 2 ($F(1, 48) = 7.62$, $p = 0.01$), suggesting learning effects or that participants changed strategies (Figure 15).

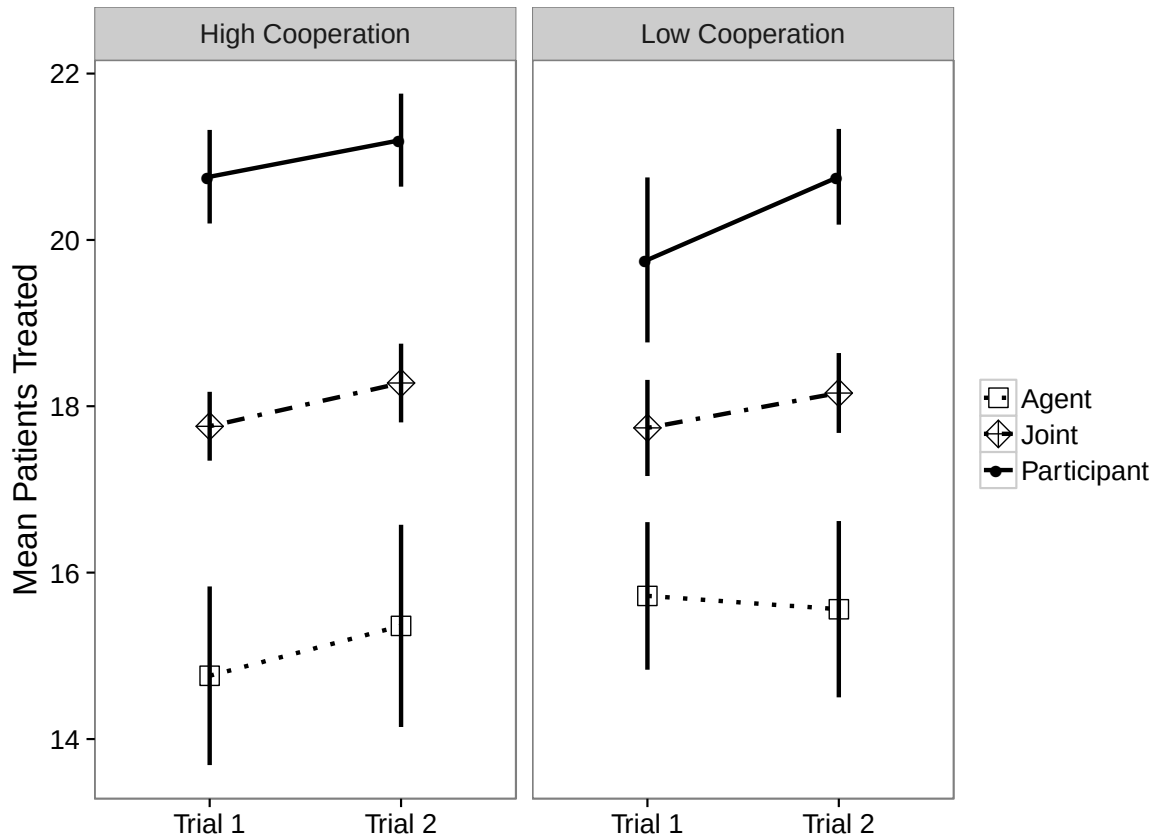

Figure 15. Comparing mean scores with 95% CIs (confidence intervals); joint score was halved for visual comparison between schedulers

Participants scored higher than agents ($F(1, 48) = 127.45$, $p < 0.01$), although this was expected. A simulation of a theoretically "perfect game" with no coordination, played by a fast (4-second assignment) and skilled player with A and B patients arriving just as the right resources become available, show a scheduler in the fast-to-slow tempo can treat a maximum of 21 patients. In the slow-to-fast tempo condition, a scheduler can treat a maximum of 16 patients. When scores were scaled to these theoretical values as percent error, no significant differences were found between conditions other than between trials. While this suggests neither group definitively benefitted from coordinating with agents, it must be taken into account that a perfect game rarely occurred, and that participants' speed varied from the 4-second pace used to calculate the theoretical perfect game. These findings also demonstrate variability in the agents' performance between trials, despite agents having relatively systematic behaviors, suggesting significant variability in people's resource exchange behaviors with the automated agents. Performance in terms of patients treated thus may not be a good indicator of cooperation and the mechanisms that lead to cooperation and resilience.

4.8.2 Resource-giving timing and utility as cooperation

Cooperative behavior was first assessed through the number of staff given in the different tempo and agent conditions. Figure 16 shows participants' and agents' giving generally corresponded to the demand patterns they experienced; each gave more during their slow-tempo period and gave less during their fast-tempo period ($F(1, 48) = 12.66$, $p < 0.01$), though tempo had a larger effect in the low-cooperation condition than in the high-cooperation condition ($F(1, 48) = 46.2$), $p < 0.01$). However, overall difference in staff given between cooperation conditions was not significant as a main effect ($F(1, 48) = 0.01$, $p = 0.92$), neither was the overall difference between trials ($F(1, 48) = 2.57$, $p = 0.12$).

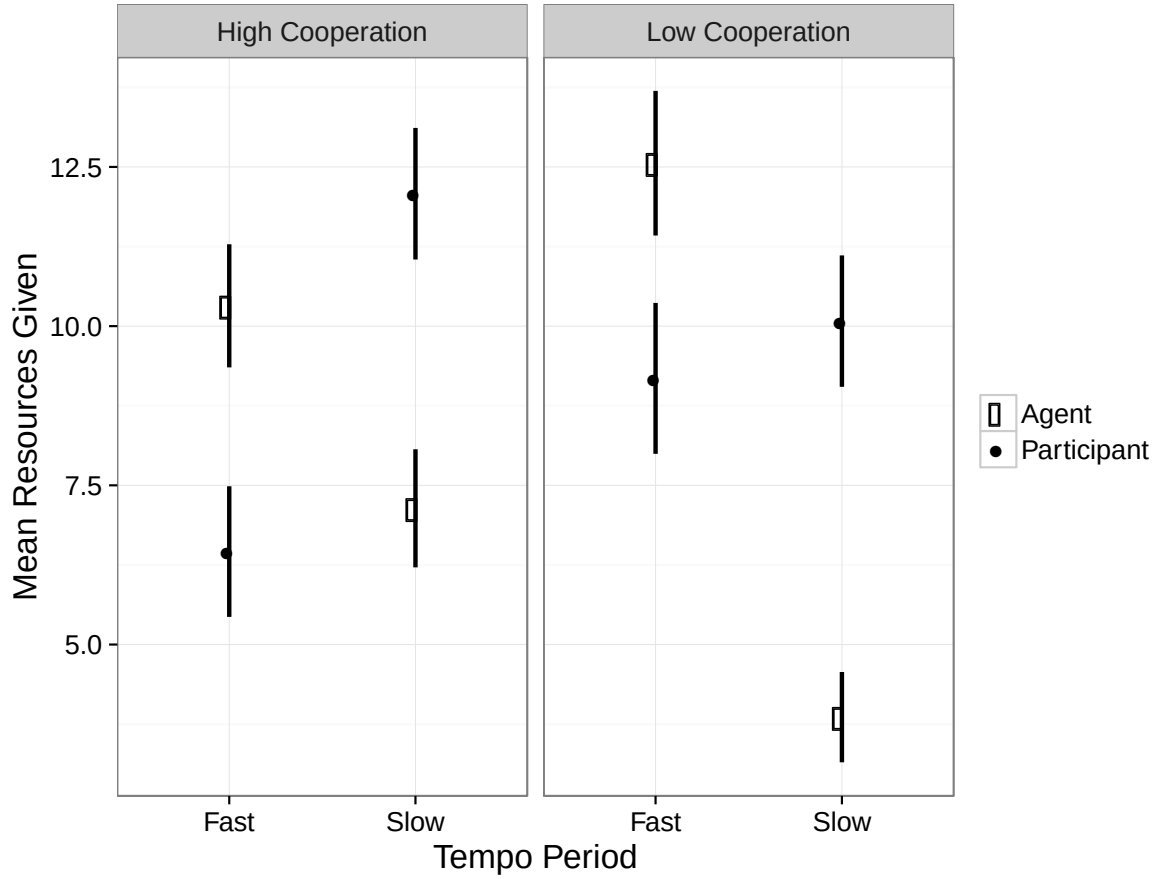

Figure 16. Staff given by cooperation conditions and tempo period with 95% CIs

Participants gave more than agents ($F(1, 48) = 64.16$, $p < 0.01$) and each gave more than the other during their slow-tempo period ($F(1, 48) = 929.68$, $p < 0.01$). In addition, more staff were exchanged during participants' fast-tempo with the low-cooperation agent ($F(1, 48) = 12.93$, $p < 0.01$); participants returned more of the staff given to them by the low-cooperation agent within the same tempo period. Such behavior indicates an effect of low-cooperation agents' indiscriminate giving of resources, and participants' inability to use many of those resources – participants returned those resources during the period when they should have withheld from giving staff to the agent.

Qualitative observations help clarify these quantitative measures of participant behavior. An important result to remark on is the low-cooperation agent seemingly engaging in more proactive behavior than the high-cooperation agent by providing more resources during participants' fast-tempo period. While this can be partially explained by the low-

cooperation agent's less discriminate giving behavior, it is also due to the interdependence of agents' behavior on participants' use of resources. While agents were designed to assign patients and staff to hospital rooms only if all resources were available, it was common for participants during experimental trials to assign patients and partial staff resources to hospital rooms prior to obtaining all staff necessary for treatment. Such a strategy of reallocating patients from the waiting room to hospital rooms with partial staff assignment may have alleviated immediate, cognitive workload in the patient assignment task, but decreased participants' spare staff pool more quickly, artificially signaling to agents that participants did not have spare staff because all were treating patients. This was not the case, as their spare staff were not treating patients but were preemptively assigned to rooms. Sensing that participants had zero of staff in their spare staff queue, this increased staff given by agents, particularly the low-cooperation agent because the situation provided more opportunities for it to provide staff indiscriminately, staff that were not as often immediately useful.

The interpretation of these observations of participant behavior influencing agent behavior that subsequently influenced participant behavior was confirmed by examining the immediate utility of resources given (Figure 17). Utility was operationalized as staff given that were used in the immediate next complete patient-resource assignment by the receiver. This calculation assumes that a resource provided that had immediate use would generally be used in the next complete staff assignment by the receiver. With this assumption, results show the high-cooperation agent gave more staff resources that were immediately used by participants than the low-cooperation agent ($F(1, 48) = 5.09$, $p = 0.03$).

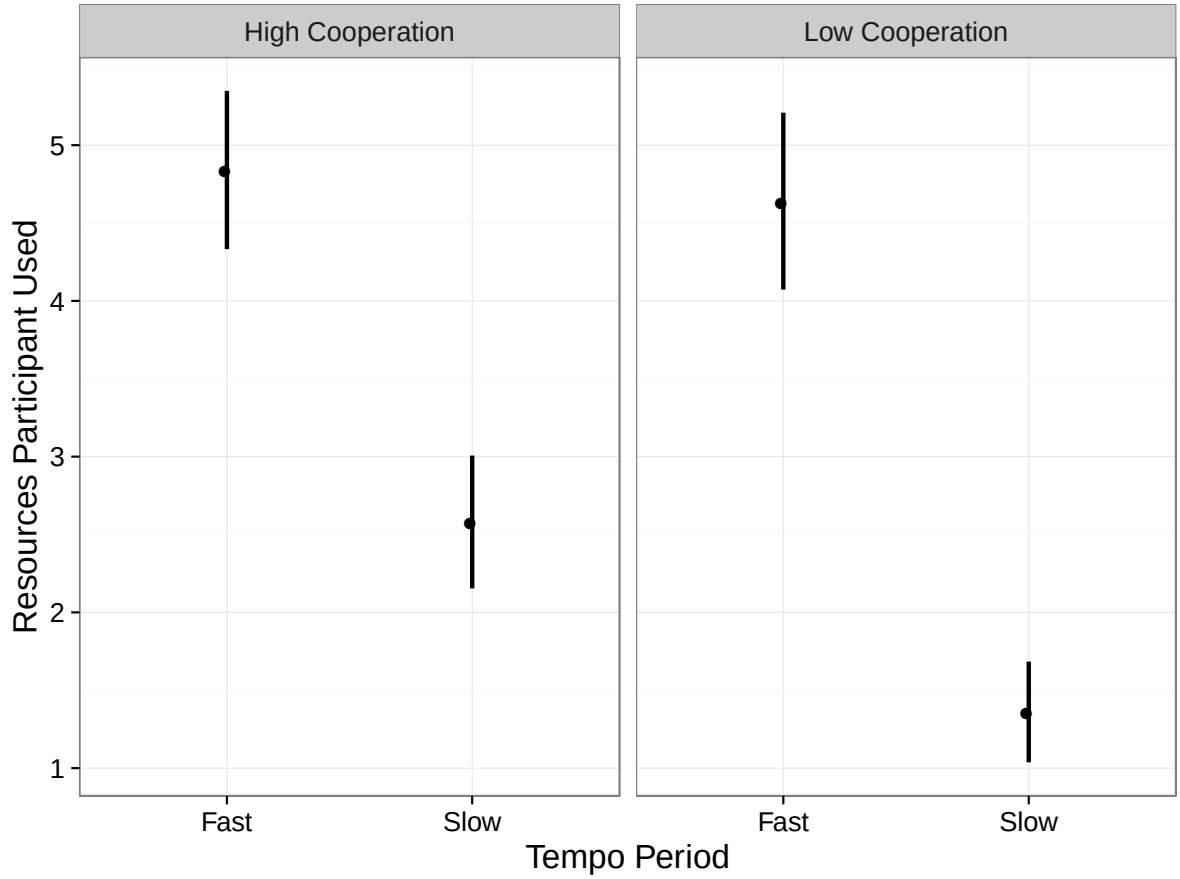

Figure 17. Number of agent staff given immediately used by participants with 95% CIs

Agents also gave more useful resources in participants' fast tempo compared to their slow tempo ($F(1, 48) = 194.66$, $p < 0.01$), with the low-cooperation agent giving fewer useful resources than the high-cooperation agent during the fast-tempo and during the slow-tempo ($F(1, 48) = 6.6$, $p = 0.01$).

Percent utility was also calculated to account for the greater number of resources provided by the low-cooperation agent during the fast-tempo period (Figure 18), by dividing the number of staff provided (with immediate utility) by the number of total staff given. While the overall difference between cooperation conditions was not significant ($F(1, 48) = 2$, $p = 0.16$), percent utility shows agents gave more staff that were immediately used by participants during participants' fast-tempo period compared to the slow-tempo period ($F(1, 48) = 4.51$, $p = 0.03$), particularly the high-cooperation agent ($F(1, 48) = 4.93$, $p = 0.03$). That participants generally used the same percent of staff given by the low-cooperation agent

across tempo conditions indicates the burden the low-cooperation enacted in creating an immediate resource-return response, rather than exchanges that led to more patients treated.

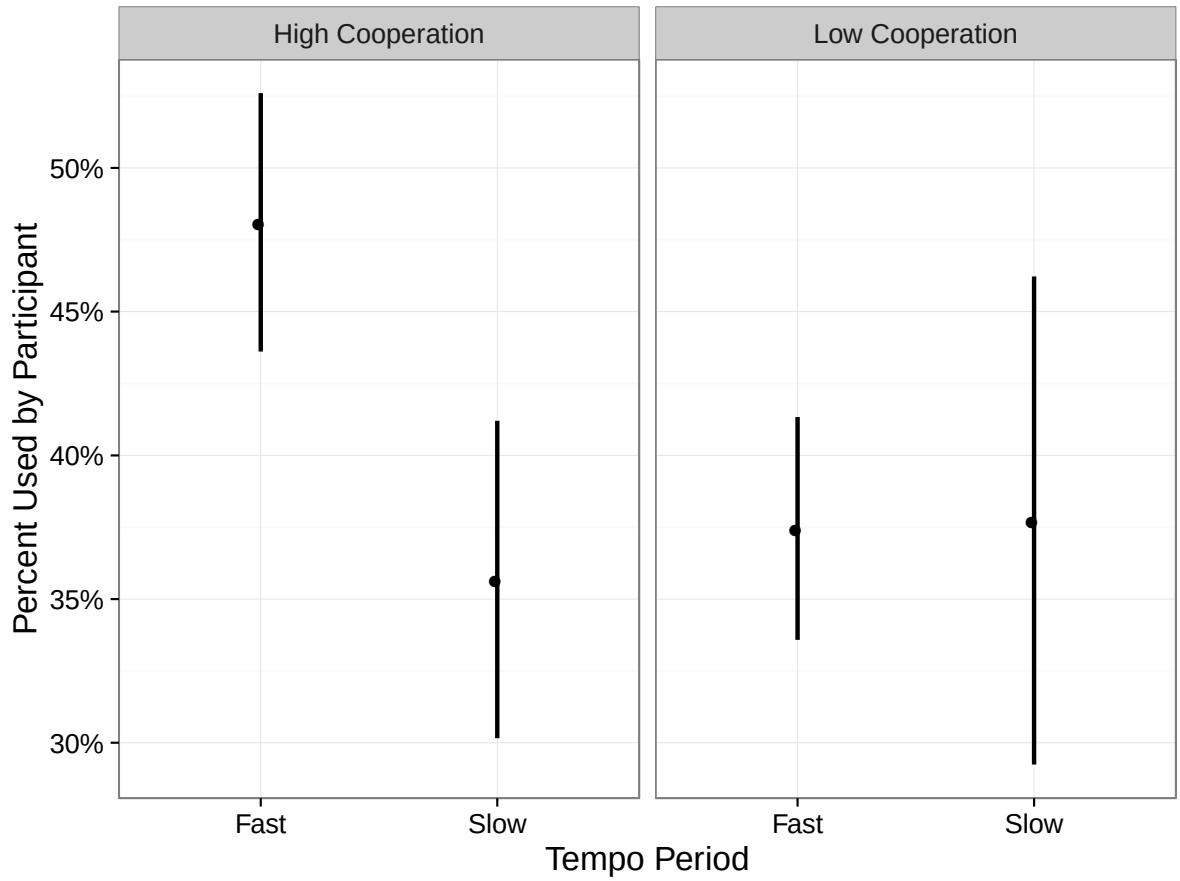

Figure 18. Percent of agent staff given immediately used by participants

The particularly wide variance in the low-cooperation, slow-tempo condition reflects several participants who used 100% staff resources provided immediately – few were provided in the first place – and several participants labeled as using 0% resources for the same reason, or because no resources were provided.

With this general understanding of what participants experienced, participants' giving behavior towards agents was examined, also by the utility of the staff they provided. Of the resources participants provided across tempo periods, more were useful to agents during participants' slow-tempo period ($F(1, 48) = 13.45$, $p < 0.01$). As Figure 19 shows, tempo period had a larger effect in the high-cooperation condition than the low-cooperation condition ($F(1, 48) = 6.29$, $p = 0.02$), meaning participants' giving behaviors were generally

more cooperative with the high-cooperation agent; they provided more useful resources when it was most appropriate to provide resources.

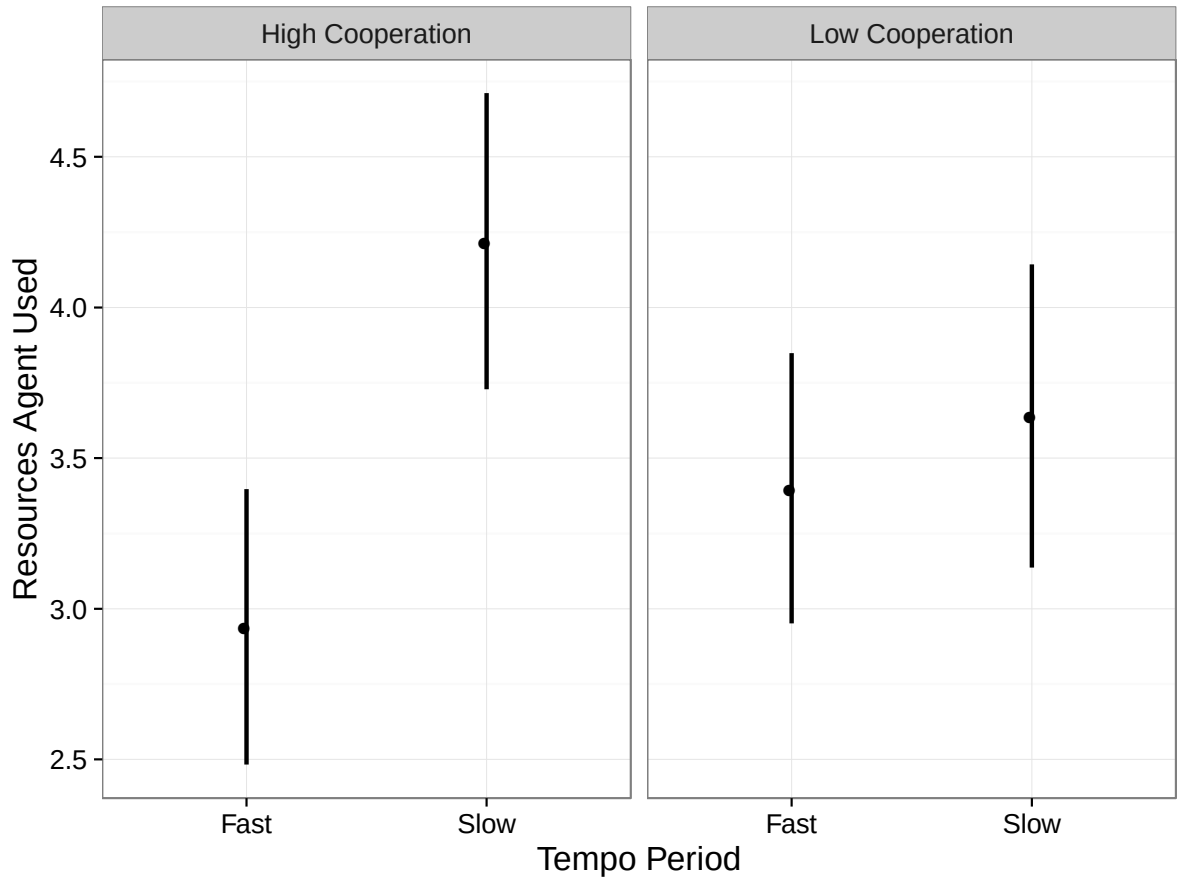

Figure 19. Number of participant staff given immediately used by agents

In the fast-tempo periods, participants providing more resources to the low-cooperation agent compared to the high-cooperation agent reflects the larger number of resources returned during this period, as shown in Figure 16. Such behaviors, returning resources while experiencing the fast-tempo, may have led to the subsequent smaller number of resources provided to the low-cooperation agent later in the trial, compared to what was provided to the high-cooperation agent. Though participants received more staff from the low-cooperation agent during their fast-tempo period, they returned fewer staff to the low-cooperation agent later in the trial. No other main effects (for cooperation ($F(1, 48) = 0.04$, $p = 0.85$) and trial ($F(1, 48) = 3.37$, $p = 0.07$)) or interaction terms were significant.

Participants' giving behavior in terms of percent utility was also calculated. Although they provided more resources to the high-cooperation agent during their slow-tempo period (Figure 19), the high-cooperation agent was not able to use many of those resources immediately (Figure 20), suggesting participants were not precise in their giving behavior, in providing immediately useful staff. In other words, participants signaled cooperation in terms of quantity of resources given, but not in terms of a deeper consideration of the type of staff agents needed.

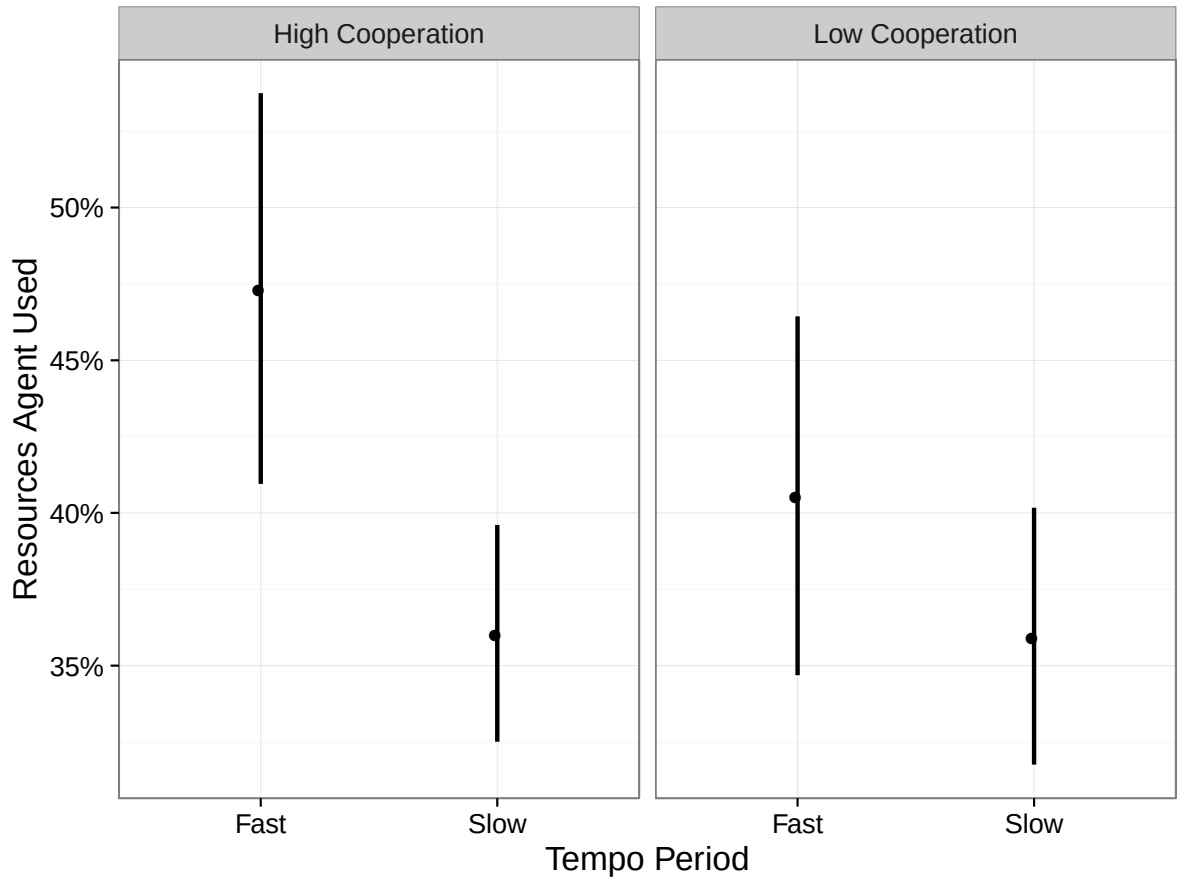

Figure 20. Percent of participant staff given immediately used by agents

Overall, a higher percent of staff participants gave were immediately used by agents during their fast-tempo period compared to the slow-tempo period ($F(1, 48) = 7.54$, $p = 0.01$). This reflects the smaller number provided to begin with (Figure 19). The high variance particularly in the fast-tempo periods reflects a few instances where agents used 100% of the few resources provided, or a few who were labeled as 0% due to no resources provided

during this period. Participants' giving behavior was therefore generally burdensome for the agents in that when they did provide more staff to the agent, those staff tended not to be immediately useful for the agent, and might have led to an eventual "return of their return" of agent-provided staff. The difference between cooperation conditions ($F(1, 48) = 1.6$, $p = 0.21$), trials ($F(1, 48) = 0.03$, $p = 0.88$), and interaction terms were not significant.

4.8.3 The role of reciprocity

The conceptualization of reciprocity in this study considers a limited, shared resource scenario in which a person with other-regarding inclinations would return the resource to the agent who, after making the first positive transfer to the participant, now has a lower margin of spare resources than participants. In a dynamic scenario like the one designed, the mere fact that the participant returned the resource to the agent would, therefore, not be evidence of positive reciprocity. This conceptualization of reciprocity allows it to be considered as a separate construct from cooperation – it is possible to demonstrate reciprocity without demonstrating cooperation in the joint goal.

Percent reciprocity was calculated by taking the percent error of participants' and agents' staff given: ((PlayerGives - AgentGives)/AgentGives). Participants with 0% reciprocity meant that by the end of the trial, an equal number of staff were returned to the agent. Results show participants interacting with the high-cooperation agent only returned about 5-6% more staff than what they received from the agent, whereas participants interacting with the low-cooperation agent returned about 13-18% more staff ($F(1, 48) = 8.31$, $p = 0.01$). Similar to the findings in Part 1, this seems to suggest that higher reciprocity with the low-cooperation agent did not necessarily result in higher joint scores.

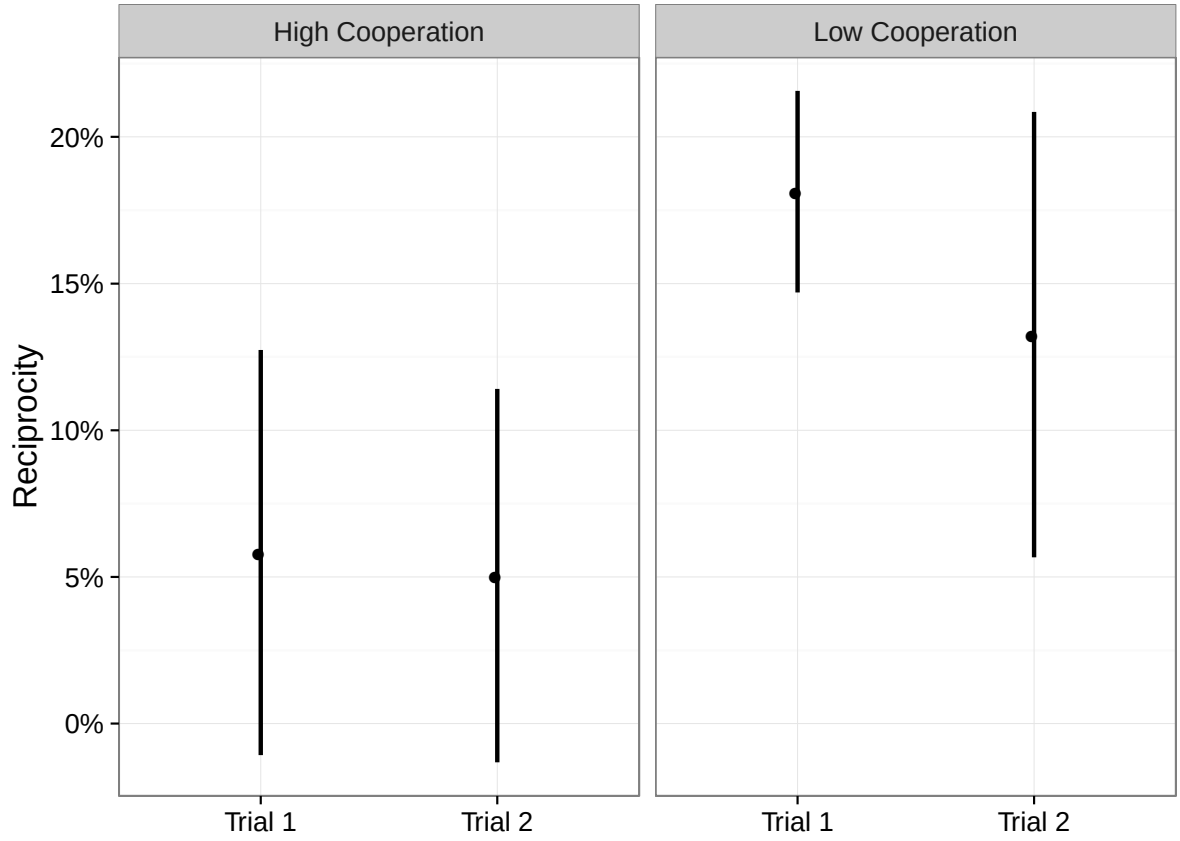

Figure 21. Participants' reciprocity per trial with the high- and low- cooperation agents

The difference between trials was not significant ($F(1, 48) = 1.37$, $p = 0.25$) nor was the interaction term for cooperation and trial ($F(1, 48) = 0.72$, $p = 0.4$). One question raised by these results is why participants had greater reciprocity with the low-cooperation condition, in that they returned more staff to the low-cooperation agent.

To examine the role of reciprocity more closely, participants were labeled as either scoring above 37 or 37 and below, a theoretical binary value for coordination benefit in the microworld. Coordination benefit is achieving greater than independent performance by working together, and the number 37 was derived from a simulated perfect game completed independently (without sharing or coordinating staff) by an "elite" human scheduler, experienced, knowledgeable, and fast with a consistent 4-second assignment time. The maximum number of patients for this elite scheduler to treat in the fast-to-slow tempo condition is 21 patients. The maximum theoretical number of patients an agent could treat in

the slow-to-fast tempo condition was 16. Therefore, the maximum theoretical joint score independently reached by elite, lucky players without coordination, would be 37. Participants who helped achieve joint scores 38 and above thus benefitted from coordination, achieving greater than the sum of individual contributions.

Participants were also labeled as having reciprocated or not, with reciprocation meaning they returned more staff than what they received in a trial. Combined with coordination benefit, results show that while participants had higher reciprocity with the low-cooperation agent, many of these participants did not achieve higher than independent joint performance (Figure 22).

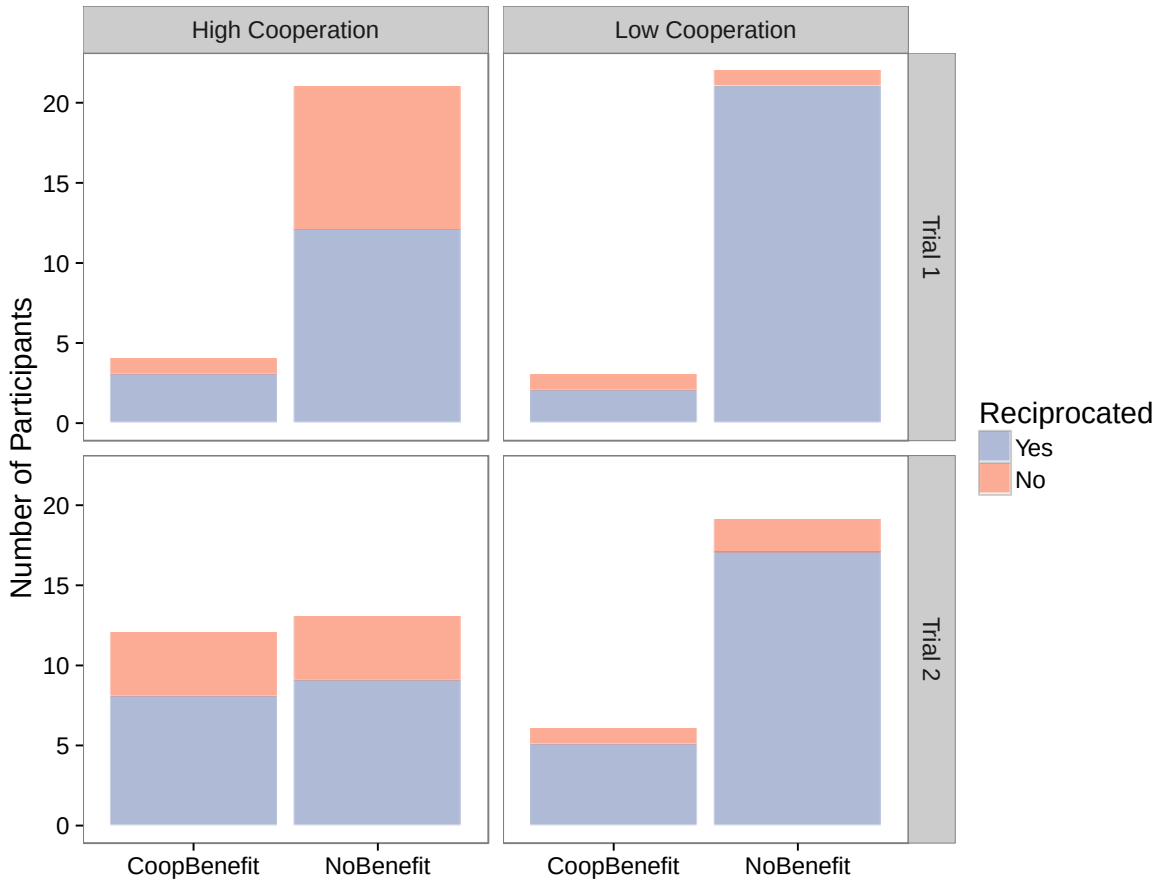

Figure 22. Labeling participants who benefitted from cooperation and reciprocated show that more reciprocity with the low-cooperation agent did not lead to better joint outcomes.

In addition, more participants in the high-cooperation condition, from reciprocating more in the second trial, were able to achieve higher scores compared to participants in the first trial.

Important to note is that in cases where slower speed and diminished skill were prominent factors in participants' performance, benefits from coordination would not be captured in this figure given the relatively high expectations of the theoretical maximum values.

4.8.4 Comparisons with Part 1, Experiment 2

To illustrate patterns of giving behavior over time and compare Part 2's reciprocal exchange to Part 1's negotiated exchange, agents' and participants' mean staff given were calculated separately per interval and visually overlaid their mean staff given per tempo. Mean differences between tempos were calculated to help explain the more detailed qualitative comparisons between intervals. Figure 23 illustrates the experience participants encountered in Part 2 – a high-cooperation agent that gave more discriminately and evenly over time, and a low-cooperation agent that gave more indiscriminately and extremely over time.

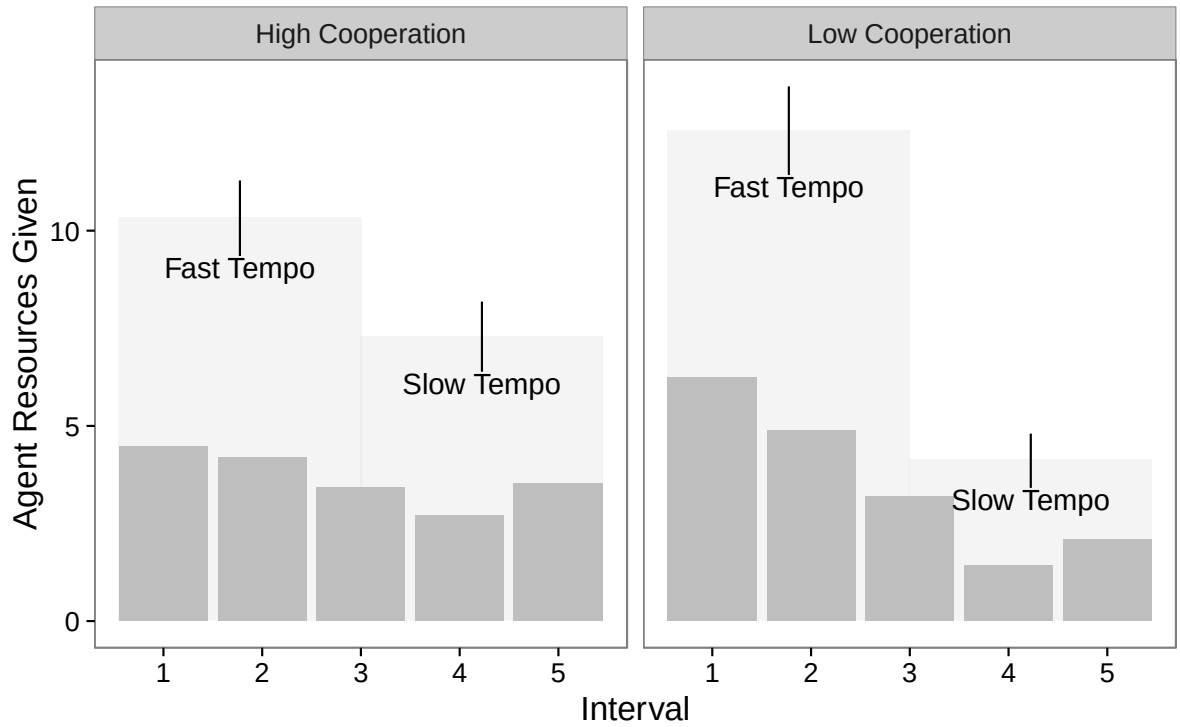

Figure 23. Part 2 reciprocal exchange agents' resource-giving patterns differed across time. Light bars are resource given per tempo with 95% CIs, dark bars are resources given per interval (for more details see section 3.3.3). Tempo is labeled from the participant's perspective.

While the overall mean staff given between cooperation conditions was not a significant difference ($F(1, 48) = 0.6$, $p = 0.44$), nor was trial ($F(1, 48) = 2.77$, $p = 0.1$), the difference between tempo was significant ($F(1, 48) = 201.65$, $p < 0.01$) as was the interaction term for tempo and cooperation ($F(1, 48) = 43.53$, $p < 0.01$). Agent giving behavior contrasted with participants' giving behavior; as shown in Figure 24, participants gave less during their fast-tempo period with the high-cooperation agent compared to with the low-cooperation agent ($F(1, 48) = 68.91$, $p < 0.01$). This supports the observation that participants' higher reciprocity with the low-cooperation agent occurred mostly during their fast-tempo period. In addition, during their slow-tempo period, participants gave more to high-cooperation agents in fast-tempo than to low-cooperation agents in fast-tempo (i.e., Figure 24, Interval 4), demonstrating more appropriate giving behavior with the high-cooperation agent ($F(1, 48) = 36.11$, $p < 0.01$).

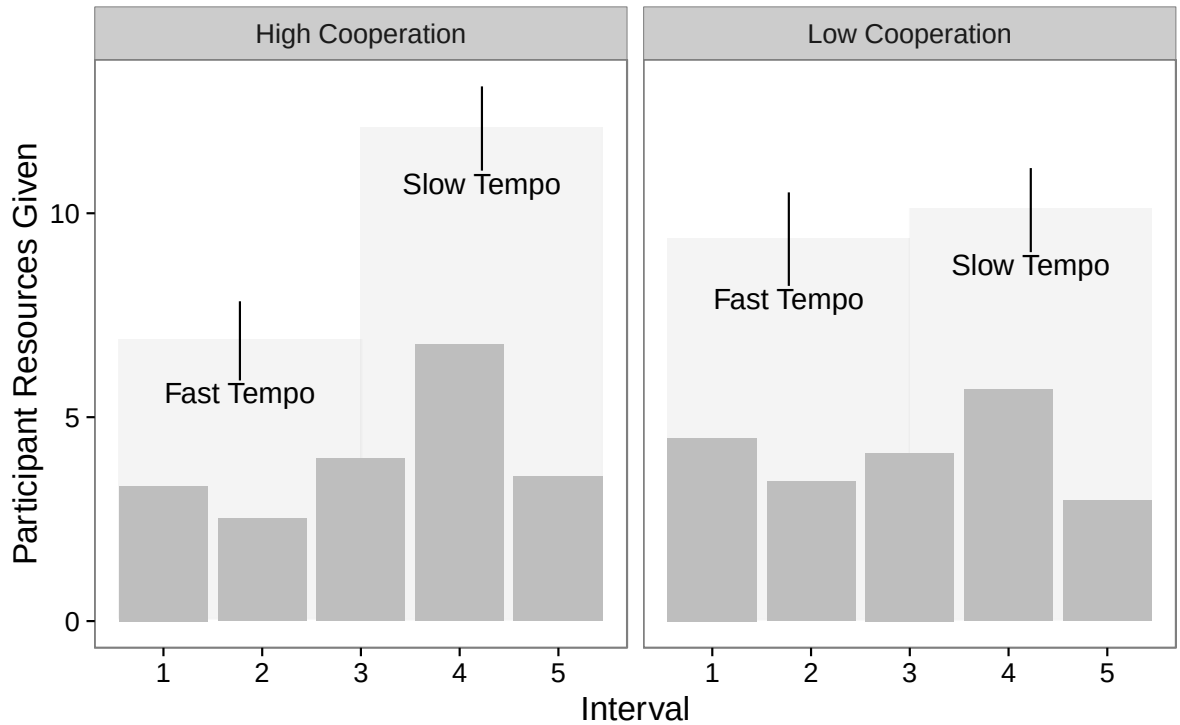

Figure 24. Part 2 reciprocal exchange participants' resource-giving behaviors

While the above interaction terms were significant, participants' overall mean staff given between cooperation conditions was not significantly different ($F(1, 48) = 0.19$, $p = 0.67$), nor was the difference between trials ($F(1, 48) = 2.22$, $p = 0.14$).

In Part 1, Experiment 2 (Chapter 3), agents' resource-giving behavior, which were accepted valid requests, was different than agents in Part 2 partly because of stronger dependence on participants' behavior, relying on participants' initiative to make requests. Overall, agents in Part 1 provided fewer staff to participants compared to agents in Part 2. While agents in Part 1 also provided differing levels of staff in the two tempo conditions ($F(1, 34) = 68.27$, $p < 0.01$), the main difference between the high- and low-cooperation agents' resource-giving behaviors is in terms of quantity, with the low-cooperation providing less staff than the high-cooperation agent (Figure 25).

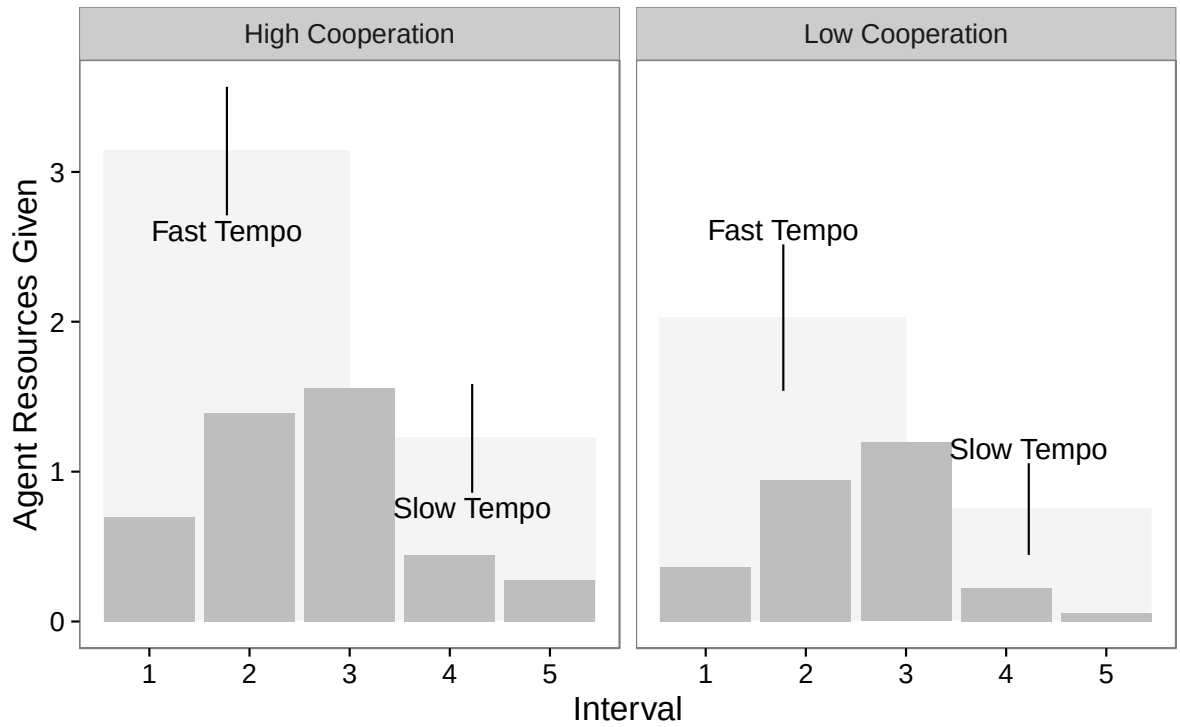

Figure 25. Part 1 negotiated exchange agents' resource-giving behaviors had similar patterns and differed mainly in quantity.

The difference between high- and low-cooperation agents in Part 2 on the other hand, is in terms of pattern over time rather than quantity. This is supported by the significant difference in mean gifts between cooperation agents in Part 1 ($F(1, 34) = 9.35$, $p < 0.01$) and the lack of

a significant difference between cooperation agents in Part 2. Differences between trials for the agents in Part 1 was not significant ($F(1, 34) = 0.83$, $p = 0.37$).

Similar differences between Part 1 and Part 2 differences were found in participants' giving behaviors; Part 1 participants generally gave differently to the high- and low-cooperation in terms of quantity ($F(1, 34) = 9.21$, $p < 0.01$) and across tempos ($F(1, 34) = 39.66$, $p < 0.01$) (Figure 26), whereas Part 2 participants' giving behaviors were different in terms of pattern rather than quantity (Figure 24). Trial also was not a significant factor in Part 1 participants' giving behaviors ($F(1, 34) = 2.3$, $p = 0.14$).

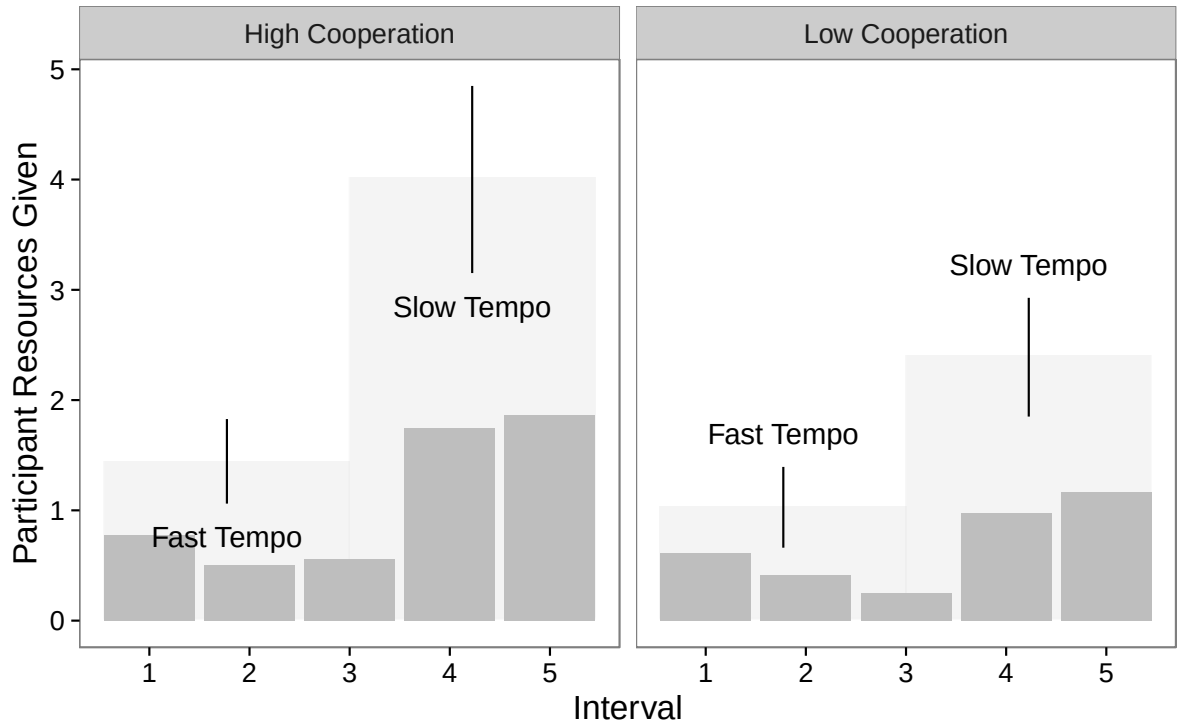

Figure 26. Part 1 negotiated exchange participants' resource-giving behaviors with the high- and low- cooperation agents also mainly differed in quantity.

It is important to remember that the sample sizes and composition for Part 1 and Part 2 participants differed, and they also had different procedures (Part 1 participants played both cooperation agents, though cooperation order was not a significant main effect). Therefore, residual variance from these factors may be a concern. However, given the similarities of the scheduling task and the unchanging hospital demand conditions in both experiments, it may

still be worth comparing how participants and agents cooperated differently in the two interaction structures, a negotiated exchange (Part 1) and a reciprocal exchange (Part 2).

Part 1 reciprocity across trials was recalculated to include only data from the first two trials in Experiment 2, which experienced the same fast-to-slow tempo as participants in Part 2. Including only the first two trials, also removed cooperation order as a factor. Plotting reciprocity from Part 1 (Experiment 2) and Part 2 together show wider confidence intervals and more variability among participants' reciprocity in Part 1 (Figure 27), again supporting the observation that dependence on participants' variable requesting behavior as well as responding behavior may have led to increased variance, compared to the relatively lower dependence on participants' behavior in Part 2 which only involved giving staff. Percent reciprocity was not significantly different between cooperation conditions in Part 1 ($F(1, 34) = 0.28$, $p = 0.6$) , and was significantly different between cooperation conditions in Part 2 $F(1, 48) = 8.31$, $p = 0.01$)

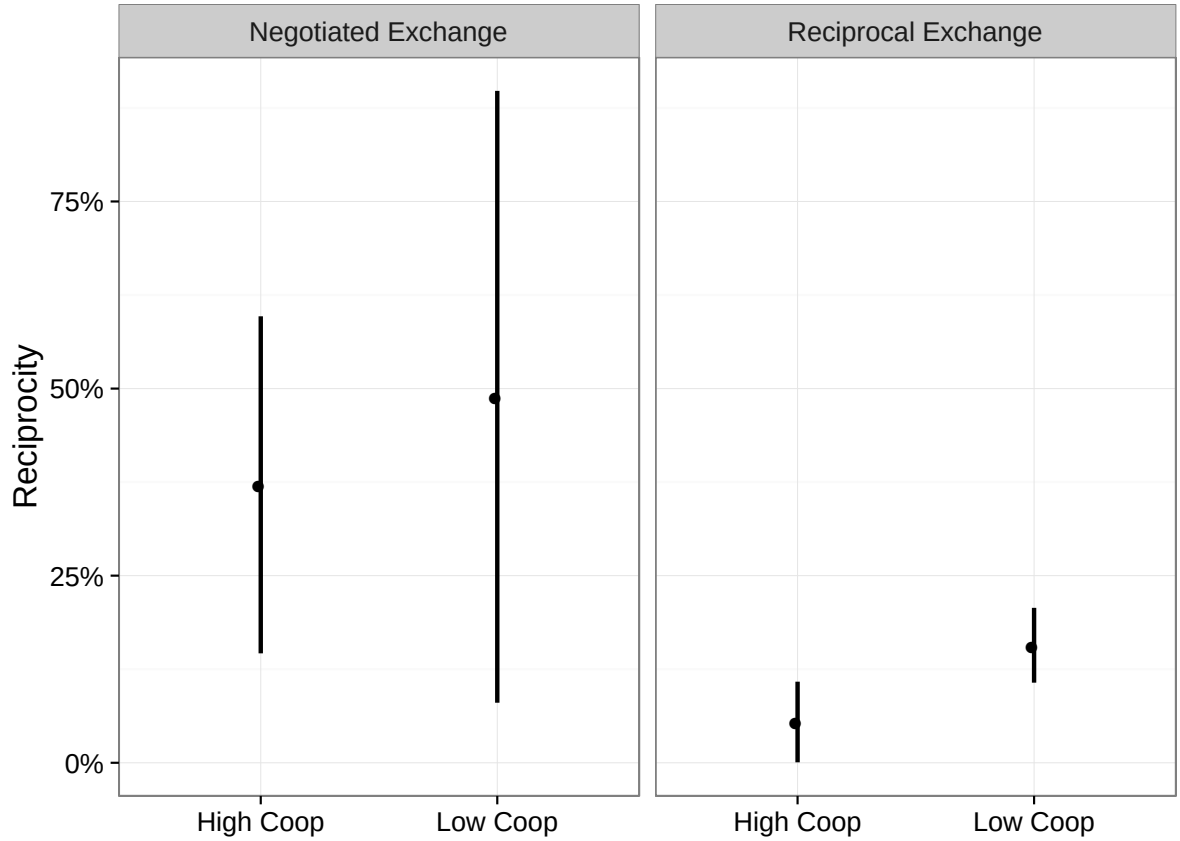

Figure 27. Comparing Part 1, Experiment 2 negotiated exchange groups (N = 18) and Part 2 reciprocal exchange groups (N = 25) shows more reciprocity and more variability in negotiated exchange.

Additionally, while there was higher reciprocity in Part 1, this was likely due to fewer staff exchanged (Figure 26), so the base threshold for matching the agents' staff was lower. In Part 2, more staff were exchanged than in Part 1 (Figure 24), and mean joint scores between the shared conditions were higher (Figure 28).

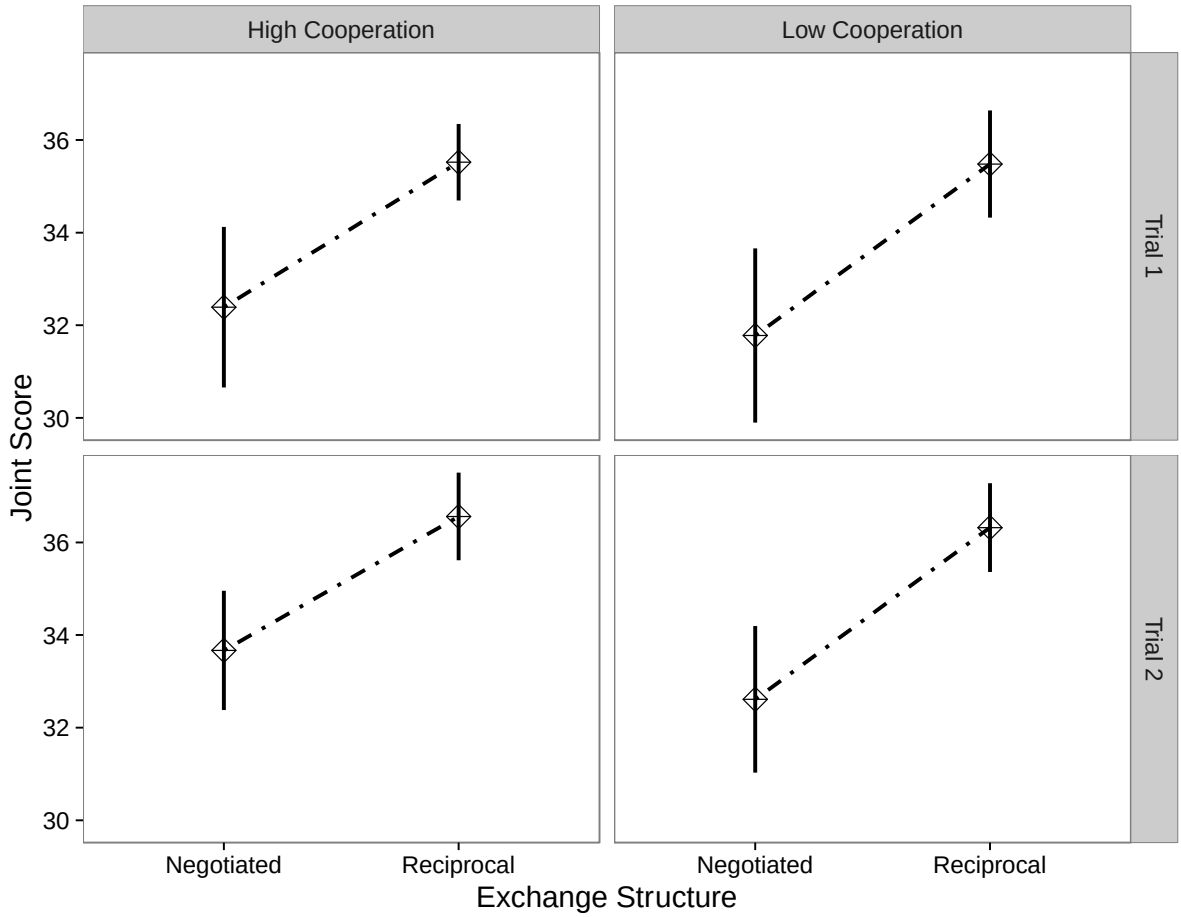

Figure 28. Joint scores across shared conditions in Part 1 and Part 2

Part 1 joint scores (M = 32.61, SD = 3.29) were lower than Part 2 joint scores (M = 35.97, SD = 2.38); $t(122.28) = -7.38$, $p < 0.01$.

4.9 General Discussion

By removing the workload needed to acquire staff through requests, Part 2 participants who passively acquired staff in the first half of the trial (during their fast tempo) were better positioned to provide the agent with staff in the second half of the trial (during their slow tempo). As expected, participants' resource provision in the second half of the trial was timelier and greater compared to Part 1. Furthermore, that people gave staff without prompting to the neighboring hospital in the second half of the trial (during participants' slow tempo) demonstrated cooperative engagement in the resource-sharing task. If workload effects dominated, participants would not have been able to give staff during the second half

of the trial when the agent needed them, and overall resource provision would have been similar to Part 1 (measured as responsive acceptances in Part 1 and proactive gifts in Part 2).

Because of changes to the task and task interface, including a shorter decision delay in the giving agent, it is difficult to say that the increased cooperation outcomes of Part 2, both in number of exchanged staff and number of treated patients, were a result of changing the social exchange structure alone. However, the improved timing of the staff exchanged during participants' slow tempo period supports the idea that a reciprocal exchange structure could be a better interaction design to support cooperation in a joint cooperation task, compared to a negotiated exchange structure. Rather than adding decision workload and relying on the agent to prompt participants of its need, as the agents did in Part 1, participants proactively considered (without prompting) the agents' needs, taking action in Part 2, and did not fall into automation-induced complacency or self-regarding tendencies by keeping all staff from the agents to themselves. The difficulty of concluding that joint performance, in terms of number of patients processed, was a function of cooperation raises the question about whether it is a useful measure for human-automation cooperation, particularly in complex, dynamic, and interdependent scenarios. Such doubt that performance in terms of number of widgets processed is not always a useful measure of important processes is reflected in the study of interactive team cognition (Cooke & Gorman, 2009), and in resilience engineering (Woods & Branlat, 2010).

Assuming a reciprocal exchange structure increased the symbolic value of interactions, compared to Part 1, it was also expected that agent cooperation level would have a greater effect on resource provision in Part 2. In other words, the differences between the high-cooperation agent and low-cooperation agent conditions would diverge more in Part 2 than in Part 1. The results did not support these expectations. The higher rate of reciprocity with the low-cooperation agent in Part 2, and the lack of significant joint score differences

suggest that although a reciprocal exchange may increase return of resources, the timing of when those resources were provided may not be appropriate for the joint goal and care should be taken in designing agent behaviors that inappropriately induce reciprocity in their human counterparts.

Participants reciprocating more with the low-cooperation agent during their fast-tempo period also suggests that an agent giving more indiscriminately but in larger quantities may be more likely to induce greater, immediate reciprocity from human counterparts, which can be counterproductive. This is particularly true in scenarios where communication between parties is limited. During a short debriefing session following the experiment with participants, an open-ended question was asked, "what did you think of the neighboring hospital agent?" About 19 participants described what was later termed as "uncooperative giving." For example, one participant partnered with the high-cooperation agent said, "I guess the agent does have the same thinking as me. When it's not busy, it sends more resources to my hospital, and if it doesn't want any of the resources it will give those back to me." Similarly, "I often felt like I was giving the agent resources it did not need and ended up giving back. Maybe because I kept giving them things I didn't need and we actually needed the same type of resource." Such feelings were not confined to participants partnered with the high-cooperation agent. From a participant partnered with the low-cooperation agent, "I think they were similar to me in the fact that we had to analyze our needs first and then contribute whatever we could to the other hospital, so that we could help the most amount of patients as possible." Such comments highlight that people generally acted less cooperatively than the high-cooperation agent, and did not recognize the high-cooperation agent's more considerate and discriminate behavior targeting participants' immediate needs.

While the high-cooperation agent took actions that were at greater risk to itself, by providing more useful resources to participants than the low-cooperation agent, it is possible

participants did not notice this while managing their own hospital. This is a critical point for understanding how signals, or lack of signals, shape people's understanding of an agent's intention. Division of labor and coordination involves personal responsibility as well as responsibility for joint outcomes within an organization. How people allocate attention and actions in service of the organizational goal while managing local goals is a question that should provide several future opportunities for research programs. The results reported in this dissertation are a first step in an empirical investigation into such a question.

Furthermore, the distinct separation of cooperation as defined by appropriate resource-sharing, and reciprocity as defined by resources returned in excess of those received, shows the complexity of each construct in more complex environments. In both exchange structures, people seemed to engage in more productive actions with the high-cooperation agent, but reciprocity as measured did not define nor inform cooperation, although it did provide a measure of how much people returned what was received. That reciprocity as measured in other studies did not explain cooperation in this study may be in part due to the "distressingly vague" longer time frames from which reciprocity seems to need to emerge (Goldstein, Pevehouse, Gerner, & Telhami, 2001, p. 597) as well as the complexity of cooperation itself, particularly in dynamic task environments, and in this study, the act of resource sharing in a fast-paced and highly interdependent task. In highly interdependent tasks, actions or inaction have a strong influence on a partner's actions or inaction, compared to less interdependent tasks (Rusbult & Van Lange, 2003). Cooperation in the scheduling task presented involved not only proactive and responsive behaviors, but also withholding of behaviors. Since withholding behaviors would be considered inaction and therefore not reciprocity, the calculation of reciprocity therefore excluded this critical component.

Withholding behaviors are typically considered part of a class of behaviors that lead to the decay of cooperation (Gächter & Herrmann, 2009). However, withholding behaviors may be cooperative, such as appropriately holding onto resources when one was given them during high demand, or appropriately refrain from requesting resources when one did not need them as badly as the agent. Such cooperative inactions echo the work on interruptions in the workplace, as discussed in section 2.4.1 – interruptions have tradeoffs. Though interruptions may be timely communication of important information, there are also cognitive costs associated with receiving and accepting an interruption. Sensitivity to such costs were incorporated into the agents' cooperativeness from the perspective of etiquette (Parasuraman & Miller, 2004); a high-cooperation agent was more considerate about a person's attention and needs, the low-cooperation agent less so.

Despite this, participants did not seem to notice or appreciate cooperative inaction, and instead responded more to a more interactive low-cooperation agent. In the negotiated exchange structure, participants returned the high number of unproductive requesting actions with the low-cooperation agent, and in the reciprocal exchange structure, unproductively returned a high number of not immediately useful resources to the low-cooperation agent during the tempo period when they needed to hold on to resources the most. This highlights perhaps some frustration involved that manifested as behavior in having limited input on a partner's decisions. More broadly speaking, the highly interdependent task and limited ability to communicate through a reciprocal exchange structure may unintentionally lead to a negotiation of sorts, i.e., communicating need through resource-giving behavior immediately following an exchange. As one participant in the reciprocal exchange group put it, "It was a burden needing to deal with someone else when you can't communicate with them."

4.10 Limitations

The question to what degree social aspects of agent cooperation influence people's cooperation, separate from task-induced cooperation, is still an open question. Other studies have explored machine behavior independent from a person's activities to measure the degree to which people responded socially (e.g., Nass & Moon, 2000; Takayama et al., 2009; Takayama, 2009). However, in most real world environments, cooperation often involves both social and economic components. It is difficult, and often not desirable, to separate the symbolic value of actions that strengthen a relationship, from instrumental value of actions that improve the immediate outcomes in joint work. Thus, while the experiments reported cannot conclude about the degree to which these differ, the novelty of this research lies in its holistic consideration of social factors in system outcomes.

4.11 Conclusion

Results from Part 2 demonstrate an overall effect of agent cooperation on human cooperation, across variation in different workload conditions and social exchange structures, with increased cooperation with a high-cooperation agent compared to a low-cooperation agent, and improved coordination in a reciprocal exchange structure compared to a negotiated exchange structure. Overall, removing control over resource acquisition and providing greater control in resource provision reduced the workload of pulling information from the agent, but also provided symbolic motivation and structural affordance for reciprocating proactive resource provision. Changing from a negotiated exchange structure to a reciprocal exchange structure on a joint coordination task not only increased resource exchange during both participants' and agents' fast-tempo periods, but also more timely resource provision.

Chapter 5 General Discussion and Conclusion

While automation is defined as "the execution by a machine agent (usually a computer) of a function that was previously carried out by a human" (Parasuraman & Riley, 1997, p. 231), it will always be a part of a larger system designed by and for people. As automation advances, understanding their changing role in human systems – and how this role affects human-automation coordination – will be key in this pursuit moving forward. For system designers to best leverage automation in increasingly dynamic and unpredictable work environments, a continuing need to understand how humans and automation can seamlessly integrate their work must consider the factors and outcomes of cooperation in service of the larger system's goals. The success of such joint systems will rely on more closely considering the human-automation relationship, including social exchange factors of their interactions. This dissertation is motivated by the need for this understanding, specifically of cooperation with an automated agent in a dynamic joint task.

A shared-resource scheduling task was developed as two microworld environments to explore human-automation cooperation, and address two main research questions: "Does agent cooperation influence people's cooperation in a joint task?" and "Does changing the interaction structure from negotiated exchange to a reciprocal exchange improve cooperation in the human-agent dyad?" Part 1 (Chapter 3) reports participants' cooperation with a high- and low- cooperation agent in a negotiated exchange structure and the influence on human-agent cooperation. Motivated by the results of unexpected workload effects in Part 1, Part 2 (Chapter 4) reports the results from two different high- and low-cooperation agents in a reciprocal exchange structure. Comparisons across Part 1 and Part 2 are reported as part of Part 2. In this chapter, the combined results are discussed more in depth, including limitations and conclusions for designers and future research.

5.1 Summary of Key Findings

Part 1 (Chapter 3) was a first step in exploring how cooperation in human-agent interactions can enhance system resilience. It looked specifically at how an automated agent's cooperation can influence people's cooperation in a dynamic joint coordination task with shared staff, and found that across variation in the task environment, people generally reciprocate higher cooperation with an agent that signals trusting, considerate behavior, and lower cooperation with an agent that signals less trusting, less considerate behavior. Part 1 also helped characterize the limitations of a negotiated exchange structure for interaction design. While negotiated exchange allows both parties to have input in decisions, it may lead to unnecessary workload, exchange delays, and disengagement (i.e., automation complacency) when it comes to considering other's needs. Having the ability to appropriately and actively consider another's needs is key to successful joint coordination, knowing when to push and pull information.

Part 2 (Chapter 4) of this dissertation, inspired by the findings from Part 1, considers an alternate interaction design, a reciprocal exchange structure involving unilateral decision-making, as a way to remove the added workload and resource acquisition delay from the resource-sharing task. Because reciprocal exchange ostensibly increases symbolic value of exchanges, Part 2 also looked at whether the effects of agent cooperation on people's cooperation were more prominent than in Part 1. Results show that although participants interacted differently with the high- and low-cooperation agents, there was greater reciprocity with the low-cooperation agent. Qualitative observation of participants' behavior partially explains this. A common strategy for participants managing demand in their hospital was to make partial patient and staff assignments to rooms, eliciting increased provision of random resources from the low-cooperation agent. Counter to expectations, this led to greater reciprocity with the low-cooperation agent during a particularly inappropriate period –

participants should have withheld providing resources to agents to better treat their high demand of incoming patients in the fast-tempo period. These results show that while people might have signaled intentions to cooperate, they failed to cooperate by returning resources to the low-cooperation agent at an inappropriate time. Thus, higher reciprocity in the coordination task was inappropriate behavior that did not lead to higher joint performance.

5.2 Cooperation to Support Resilience

Investigating human-agent cooperation in a dynamic joint task contributes to Resilience Engineering because of the important role cooperation plays among networked people and agents, interacting and coordinating in an ad hoc manner to address unexpected demands as they occur. In such environments, where coordination is difficult to preplan, cooperation is needed. When people are involved, social processes facilitate cooperation. Understanding how people cooperate with increasingly autonomous automation will be important for avoiding automation designs that elicit or encourage unproductive behaviors from their human counterparts. Likewise, better understanding these social processes and the structures that influence them may improve future automation design and their integration in human systems.

This work approaches cooperation by considering cooperation behaviors (signals and actions), cooperation processes, and cooperation outcomes. This multi-tiered approach adds to a better understanding of how micro-level behaviors lead to processes and macro-level outcomes. In using this approach to investigating cooperation, a more holistic assessment was possible, which emphasizes the conclusion that cooperation is a nuanced construct that depends on its context, and cannot be defined as simply providing more resources. Rather, providing resources cooperatively requires not only the willingness to sacrifice local resources for the shared good, but also consideration of timing and need in providing those resources. Furthermore, while joint outcome and other summative measures can be useful as

measures of joint performance, they may not be enough to understand the underlying processes leading to joint performance.

5.3 Limitations

The results from this research have important implications for human-automation interaction, particularly increasingly autonomous agents interacting dynamically with human counterparts. Though inferences can be made about the effects of agent behavior and the interaction structure on people's cooperative behaviors, attempting to address the degree to which a comprehensive range of factors influence human-automation interaction would be unwieldy given the intended focus of this dissertation. The results must therefore be considered through the lens afforded by the research objectives, the population sample tested, and the microworld environment.

The microworld environment developed for this research was not intended to mimic a real world hospital scheduling scenario, but more simply a generic cooperation situation that could be completed within an hour for practical and convenience reasons. It should be noted that the outcomes of this research may differ depending the context of the work, and the level of risk involved in the individual task and in the shared task. In general, the population sample tested was biased and not random, as it comprised a generally younger adult population living or working near a university community in a developed country. While this allows more straightforward comparison with studies that have similar demographics, it contributes little to the overarching knowledge base of human cooperation or larger group cooperation, including potential differences between cultures (Fehr & Fischbacher, 2003). The reasons for the studies' time limit were both budgetary and to avoid participant fatigue and disengagement (Cummings, 2015). Groups of people who are required to work longer periods of time with automation may be more susceptible to fatigue and its associated negative effects (Barker & Nussbaum, 2011; Smith, Carayon, Sanders, Lim, & Legrande,

1992), and specialized groups that train with automation in actual work scenarios may have different mental models or attitudes toward automation that would affect cooperation and that differ from a more general population.

Furthermore, this dissertation avoids the ongoing discussion of differences between interactions with other people and interactions with machines. Previous research reports conflicting findings. Some studies report differences in how people treat other people compared to how they treat machines, and rightly caution against generalizing between the two (Demir & Cooke, 2014; Groom & Nass, 2007; Lee & See, 2004; Madhavan & Wiegmann, 2007; Mcknight, Carter, Thatcher, & Clay, 2011). Other studies report a lack of difference in certain circumstances, or suggest compelling similarities in the way people behave with machines and the way they behave with people (Fogg & Nass, 1997; McCabe, Houser, Ryan, Smith, & Trouard, 2001; Nass & Moon, 2000; Parasuraman & Miller, 2004). Trends in robotics certainly seem to be taking directions that make machines more like people (Breazeal, 2000; Kaplan, 2001; Mutlu et al., 2012); however, differences in the effects of perceived agency and perceived competence on cooperation may influence how people respond to machines. This dissertation does not speculate on how participants may or may not have treated the agent were they told it was another person. What is important for automation designers is to be aware of potential differences, and to test and evaluate these factors. For researchers, they must be sensitive to potential differences when generalizing from exchange studies between people and exchanges between people and machines.

5.4 Suggestions for Practitioners

When designing direct exchange between people and automated agents, this study found that a negotiated exchange structure allowed for better communication of need, affording more accountability for each individual scheduler's acquisition of resources. However, this led to cognitive tunneling among participants who experienced an initial high

demand, encouraged a more autonomous strategy, and a failure to request enough resources during the high demand period. A potential solution to help reduce individual cognitive tunneling, the reciprocal exchange structure allowed schedulers to make decisions without needing to check in with one another. This led to increased resource exchanges and higher joint scores, though participants lacked sensitivity of the agent's status or need, particularly with the more insensitive and more active (low cooperation) agent. Though the two exchange structures were tested separately to determine their differences, a mix of both interactions structures that allow participants and agents to both push and pull the right information at the right time may provide a better outcome, depending on the task context.

In terms of agent cooperation, it is important to note that a highly cooperative agent that is sensitive to others' needs may go unnoticed if its sensitivity is expressed as inaction, i.e., not taking an inappropriate action. In Part 2, we found participants were insensitive to both sensitive and insensitive agents, and additionally, reflected the more insensitive agent's behavior by providing resources to the insensitive agent in response to its behavior rather than in response to environment and individual and global need. Thus, agent cooperation should be additionally considered in the context of the social exchange structure. Since participants were unable to negotiate with the agent on what resources they needed at a particular time, they may have tried to use the only action they had – resource provision – in their attempts to better coordinate with the agent.

5.5 Suggestions for Future Research

This work demonstrated two ways to explore cooperation in human-automation joint tasks — through agent behavior and the interaction structure. However, these are just two of many possible lines of inquiry, a few of which are expanded on below, derived from the findings of this dissertation.

5.5.1 Cooperation when goals conflict

In Part 2's reciprocal exchange structure, where decisions were unilateral, results show that indiscriminately giving in large quantities, despite potentially negative effects on the joint task, may have led to inappropriate resource provision. Cooperation is not simply providing resources, but also whether that provision fits in the context of the joint task environment and shifting priorities, which the measure of reciprocity used did not capture on its own. Another way to measure cooperation could be the degree to which people are willing to self-sacrifice for the greater good, e.g., actions taken when goals inadvertently conflict, or actions taken when local margins are depleted, rather than in excess, as was implemented in this research. An alternate microworld where demand patterns conflict rather than complement may yield interesting insights for this question.

5.5.2 Collegiality and competence

The term *automation collegiality* was introduced at the beginning of this work to describe a relationship between peers, or laterally organized entities, who have a common goal. Whereas cooperation was defined as self-sacrifice in service of this common goal, the collegiality of an entity would entail its signals of intention, as well as the perception of its intentions. While acts, signals, and perception are not exclusive, separating the act of self-sacrifice from signals (and perception) of intention allows a discussion on how signals of cooperation might differ from acts of cooperation. This dissertation mainly focuses on agents' signals of cooperation without optimizing for actual cooperation given the known demand patterns designed by the researcher.

Past research in human-automation interaction tends to focus on how the joint human-automation system varies with automation performance; this continues to be important. While not the focus of this dissertation, perceptions of the automation's capability can still influence how people choose to interact with the automation, even in more lateral

coordination scenarios like the one studied. During a short debriefing session of the experiment with participants, an open-ended question was asked, "what did you think of the neighboring hospital agent?" Around 13 participants mentioned agent performance related to its capability. One participant said, paraphrased, "I'm not sure whether the neighboring hospital agent is optimizing its strategy. That's why I put priority in treating my own patients even when I saw the other hospital was red. I was worried that even if I sent resources to the neighboring hospital, it would not use them correctly."

This suggests there is an additional gap in the research, the relationship between collegiality and competence. The extent to which collegiality and competence are related is still unknown, let alone how different situations might mediate their relationship. Some situations might benefit more from automation collegiality where automation competence is difficult and the consequences are low (Takayama, Dooley, & Ju, 2011); situations with higher risk may not benefit as much. Future work on collegiality and competence could look more closely at the effects that the intention to fulfill requests, apart from fulfilling requests appropriately, might have on human-agent cooperation. For example, do acts signaling intention to cooperate without actual instrumental value facilitate a cooperative relationship, or do they undermine perceptions of capability and subsequently undermine cooperation?

5.5.3 Cognitive effort as resources in joint action

This research focuses on previously neglected factors in joint human-automation system performance by taking a more relational approach. By applying social exchange theory, much was gleaned from studying human-automation cooperation in the context of a dynamic task with limited resources. Most notably, this dissertation raises an aspect of cooperation not often addressed by more theoretical accounts (Axelrod, 1984; Castelfranchi & Falcone, 2001; Hoc, 2000, 2013), including the investment of cognitive resources to facilitate joint action. Whereas game-theoretic accounts establish cooperation as a tradeoff of

instrumental risks and benefits, and social-psychological accounts establish cooperation as affective or symbolic processes in social relationships, the human-factors account from this dissertation contributes the idea that cooperation involves not only the sharing of material resources, but also the sharing of cognitive resources, and that the work context matters. Cooperation not only involves tradeoffs from exchanging instrumental resources, but also tradeoffs from exchanging cognitive resources, such as putting effort into the timing of when to give instrumental resources to minimize the negative effects of interruptions (McFarlane & Latorella, 2002), or of offloading decision-making workload. Taking a social exchange approach to cognitive resources can apply to the study of interruptions, as a way to structure a study weighing the benefits and advantages of interruptions in time-critical work environments. For example, do the costs of interruptions outweigh the benefits (Grundgeiger & Sanderson, 2009), and how might such costs be reduced through cooperation and other-regarding behavior?

5.5.4 Technology-mediated cooperation between people

While this dissertation focused specifically on human-automation cooperation, parallels may be drawn to research in technology-mediated human-human cooperation. As ubiquity of information and communication technologies virtualize work environments, insights from studies on interactions with increasingly autonomous agents may both derive inspiration from and inspire research on interactions between people in virtual environments. Future work in distinguishing the differences between these two types of partners may explore the degree to which sensitivity to automation capability and error, compared to human capability and error, plays a role in lateral control contexts. People's tendency to overestimate the role of dispositional factors and underestimate the role of environmental factors in automation (Madhavan, Wiegmann, & Lacson, 2006; Muir, 1987) may lead to certain patterns of cooperation depending on agent signaling and information about the

environment in a display interface. During the open-ended debriefing of Part 2, in which participants were asked, “what did you think of the neighboring hospital agent?” 8 participants mentioned environmental factors influencing the agent and giving dynamics, while 9 mentioned the agent’s intention. Future studies may want to further explore the extent to which people are sensitive to an automated agent’s capability compared to human compatibility in more lateral control situations.

5.6 Conclusion and Contributions

This dissertation presents among the first known empirical studies to explicitly consider human-automation interaction in a dynamic task environment from the perspective of social exchange behaviors with more autonomous agents. What prior human-automation interaction research neglects is an investigation into how social exchange factors of cooperation contribute to resilience. This dissertation broached this gap by investigating human-automation cooperation in a dynamic shared-resource task. Furthermore, prior social exchange research does not include the idea that cognitive effort is also a resource that is invested to facilitate joint action. The results presented showed cooperation with an automated agent must include consideration of this type of cognitive investment, as it may influence cooperation apart from the instrumental or symbolic value of an exchange. Future work will need to explore this relationship more fully. The main contributions of this work thus center around insights into cooperation and reciprocity in human-automation interaction, and how research in this area might be conducted to support the design of resilience in future systems. Findings from this research can help designers of such systems identify and avoid potential coordination breakdowns in dynamic joint task environments. Most practically, this work demonstrates that factors influencing human-agent cooperation include signals of cooperative behavior and the social exchange structure of the interaction.

References

Adams, B., Breazeal, C., Brooks, R., & Scassellati, B. (2000). Humanoid robots: A new kind of tool. *IEEE Intelligent Systems and Their Applications*, *15*(4), 25–31. doi:10.1037/e448722006-001

Adler, P. (1997). Work organization: From Taylorism to teamwork by. *Perspectives on Work*, (June), 61–65.

Adler, P. (2001). Market, hierarchy, and trust: The knowledge economy and the future of capitalism. *Organization Science*, *12*(2), 215–234.

Allen, J., Guinn, C., & Horvitz, E. (1999). Mixed-initiative interaction. *IEEE Intelligent Systems and Their Applications*, *14*(5), 14–23. doi:10.1109/5254.796083

Ariely, D. (2008). *Predictably Irrational*. New York, NY: Harper-Collins Publishers.

Arrow, K. (1974). *The Limits of Organizations* (First.). New York: Fels Center of Government.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–6.

Bailey, B., & Iqbal, S. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction*, *14*(4), 1–28. doi:10.1145/1314683.1314689

Bailey, B., & Konstan, J. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, *22*(4), 685–708. doi:10.1016/j.chb.2005.12.009

Bailey, N., & Scerbo, M. (2007). Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science*, *8*(4), 321–348. doi:10.1080/14639220500535301

Baker, D., Day, R., & Salas, E. (2006). Teamwork as an essential component of high-reliability organizations. *Health Services Research*, *41*(4 Pt 2), 1576–98. doi:10.1111/j.1475-6773.2006.00566.x

Barker, L., & Nussbaum, M. (2011). Fatigue, performance and the work environment: a survey of registered nurses. *Journal of Advanced Nursing*, *67*(6), 1370–82. doi:10.1111/j.1365-2648.2010.05597.x

Barnes, M., Chen, J., Jentsch, F., Oron-Gilad, T., Redden, E., Elliott, L., & Evans III, A. (2014). Designing for Humans in Autonomous Systems: Military Applications, (January 2014).

Barrett, L., Gaynor, D., & Henzi, S. (2002). A dynamic interaction between aggression and grooming reciprocity among female chacma baboons. *Animal Behaviour*, *63*, 1047–1053. doi:10.1006/anbe.2002.3008

Benninghoff, B., Kulms, P., Hoffmann, L., & Krämer, N. (2012). Theory of mind in human-robot-communication: Appreciated or not? In A. Kluge & R. Söffker (Eds.), *Inter-*

disziplinärer Workshop Kognitive Sys-teme: Mensch, Teams, Systeme und Automaten (2nd ed.).

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*, 122–142.

Bertuccelli, L., & Cummings, M. (2011). Scenario-based robust scheduling for collaborative human-UAV visual search tasks. In *50th IEEE Conference on Decision and Control and European Control Conference* (pp. 5702–5707). Orlando, FL.

Billinton, R., & Allan, R. (1983). Introduction. In *Reliability Evaluation of Engineering Systems: Concepts and Techniques* (pp. 1–4). Springer US. doi:10.1007/978-1-4615-7728-7_1

Bishop, J., & Scott, K. (2000). An examination of organizational and team commitment in a self-directed team environment. *The Journal of Applied Psychology*, *85*(3), 439–450. doi:10.1037/0021-9010.85.3.439

Blomqvist, K., & Stahle, P. (2004). Trust in technology partnerships. In M. Huotari & M. Ivonen (Eds.), *Trust in Knowledge Management and Systems in Organizations* (pp. 173–199). Hershey, PA: Idea Group Publishing. doi:10.4018/978-1-59140-126-1.ch008

Bottom, W., Holloway, J., Miller, G., Mislin, A., & Whitford, A. (2006). Building a pathway to cooperation: Negotiation and social exchange between principal and agent. *Administrative Science Quarterly*, *51*, 29–58. doi:10.2189/asqu.51.1.29

Bowling, M., Burch, N., Johanson, M., & Tammelin, O. (2015). Heads-up limit hold ’ em poker is solved. *Science*, *347*(6218), 145–149.

Bradshaw, J., Feltovich, P., Jung, H., Kulkarni, S., Taysom, W., & Uszok, A. (2004). Dimensions of adjustable autonomy and mixed-initiative interaction. In M. Nickles, M. Rovatsos, & G. Weiss (Eds.), *Agents and Computational Autonomy: Potential, Risks, and Solutions. Lecture Notes in Computer Science* (pp. 17–39). Berlin: Springer Verlag.

Bray, L., Anumandla, S., & Thibeault, C. (2012). Real-time human–robot interaction underlying neurorobotic trust and intent recognition. *Neural Networks*.

Breazeal, C. (2000). *Sociable machines: Expressive social exchange between humans and robots (Doctoral dissertation)*. Massachusetts Institute of Technology.

Breazeal, C., & Scassellati, B. (1999). How to build robots that make friends and influence people. *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No.99CH36289)*, *2*, 858–863. doi:10.1109/IROS.1999.812787

Brooks, R. (1991). Intelligence without reason. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence* (pp. 569–595). Morgan Kaufmann.

Casper, J., & Murphy, R. R. (2003). Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Transactions on Systems, Man, and Cybernetics. Part B: Cybernetics*, *33*(3), 367–85. doi:10.1109/TSMCB.2003.811794

Cassell, J., & Bickmore, T. (2000). External manifestations of trustworthiness in the interface. *Communications of the ACM*, *43*(12), 50–56.

Castelfranchi, C. (1998). Modelling social action for AI agents. *Artificial Intelligence*, *103*, 157–182.

Castelfranchi, C., & Falcone, R. (2001). Social trust: A cognitive approach. In *Trust and Deception in Virtual Societies* (pp. 55–90). Citeseer.

Chiou, E., & Lee, J. (2016). Cooperation in human-agent systems to support resilience: A microworld experiment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. doi:10.1177/0018720816649094

Clark, H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Cooke, N., & Gorman, J. (2009). Interaction-Based Measures of Cognitive Systems. *Journal of Cognitive Engineering and Decision Making*, *3*(1), 27–46. doi:10.1518/155534309X433302

Cooke, N., Gorman, J., Myers, C., & Duran, J. (2013). Interactive team cognition. *Cognitive Science*, *37*(2), 255–85. doi:10.1111/cogs.12009

Cox, J. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*(2), 260–281. doi:10.1016/S0899-8256(03)00119-2

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, M., … Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496.

Cropanzano, R., & Mitchell, M. (2005). Social exchange theory: An interdisciplinary review. *Journal of Management*, *31*(6), 874–900. doi:10.1177/0149206305279602

Cummings, M. (2015). Boredom in the workplace: A new look at an old problem. *Human Factors*, *1*, 1–30. doi:10.1017/CBO9781107415324.004

Cummings, T. (1978). Self-regulating work groups: A socio-technical synthesis. *The Academy of Management Review*, *3*(3), 625–634.

Czerwinski, M., Cutrell, E., & Horvitz, E. (2000). Instant messaging and interruption: Influence of task type on performance. *Proceedings of OZCHI 2000*, 356–361. doi:10.1016/S1361-3723(02)01112-0

Dabbish, L., & Kraut, R. (2004). Controlling interruptions : Awareness displays and social motivation for coordination. In *Proceedings of CSCW'14* (pp. 182–191). Chicago: ACM Press.

De Dreu, C., Greer, L., Handgraaf, M., Shalvi, S., Van Kleef, G., Baas, M., … Feith, S. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science*, *328*(5984), 1408–1411. doi:10.1126/science.1189047

de Visser, E., & Parasuraman, R. (2011). Adaptive Aiding of Human-Robot Teaming: Effects of Imperfect Automation on Performance, Trust, and Workload. *Journal of Cognitive Engineering and Decision Making*, *5*(2), 209–231. doi:10.1177/1555343411410160

Defense Science Board Task Force. (2012). *The Role of Autonomy in DoD Systems*. Washington, D.C.

Delton, A., Krasnow, M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(32), 13335–13340. doi:10.1073/pnas.1102131108

Demir, M., & Cooke, N. (2014). Human teaming changes driven by expectations of a synthetic teammate. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 16–20. doi:10.1177/1541931214581004

Desteno, D., Breazeal, C., Frank, R., Pizarro, D., Baumann, J., Dickens, L., & Lee, J. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, *23*(12), 1549–56. doi:10.1177/0956797612448793

Dugatkin, L., Mesterton-Gibbons, M., & Houston, A. (1992). Beyond the prisoner's dilemma: Toward models to discriminate among mechanisms of cooperation in nature. *Trends in Ecology & Evolution*, *7*(6), 202–205.

Endsley, M. (1996). Automation and situation awareness. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 163–181). Mahwah, NJ: Lawrence Erlbaum.

Endsley, M., & Kaber, D. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492.

Epley, N., Waytz, A., & Cacioppo, J. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. doi:10.1037/0033-295X.114.4.864

Evans, A., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, *42*(6), 1585–1593. doi:10.1016/j.jrp.2008.07.011

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–91. doi:10.1038/nature02043

Fehr, E., & Gächter, S. (1998). Reciprocity and economics: The economic implications of Homo Reciprocans. *European Economic Review*, *42*(3-5), 845–859. doi:10.1016/S0014-2921(97)00131-1

Ferrin, D., Bligh, M., & Kohles, J. (2007). Can I trust you to trust me? A theory of trust, monitoring, and cooperation in interpersonal and intergroup relationships. *Group & Organization Management*, *32*(4), 465–499.

Ferrucci, D. (2012). Introduction to "this is watson." *IBM Journal of Research and Development*, *56*(3.4).

Fink, R., & Weyer, J. (2014). Interaction of human actors and non-human agents: A sociological simulation model of hybrid systems. *Science, Technology & Innovation Studies*, *10*(1), 47–64.

Fitts, P., Viteles, M., Barr, N., Brimhall, D., Finch, G., Gardner, E., … Stevens, S. (1951). *Human engineering for an effective air-navigation and traffic-control system*. Washington, D.C.: National Research Council.

Fogg, B., & Nass, C. (1997). How Users Reciprocate to Computers: An experiment that demonstrates behavior change. In *CHI '97* (pp. 331–332).

Fong, T., Nourbakhsh, I., Kunz, C., Fluckiger, L., Schreiner, J., Ambrose, R., … Scholtz, J. (2005). The peer-to-peer human-robot interaction project. In *Proceedings from SPACE Conferences and Exposition: Space 2005*. Reston, VA: American Institute of Aeronautics and Astronautics. doi:10.2514/6.2005-6750

Franklin, S., & Graesser, A. (1997). Is it an agent, or just a program?: A taxonomy for

autonomous agents. *ECAI '96: Proceedings of the Workshop on Intelligent Agents III, Agent Theories, Architectures, and Languages*, 21–35. doi:10.1007/BFb0013570

Frey, C., & Osborne, M. (2013). *The future of employment: how susceptible are jobs to computerisation?*

Fussell, S., Kiesler, S., Setlock, L., & Yew, V. (2008). How people anthropomorphize robots. *Proceedings of the 3rd International Conference on Human Robot Interaction - HRI '08*, (c), 145. doi:10.1145/1349822.1349842

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1518), 791–806. doi:10.1098/rstb.2008.0275

Gao, J., Lee, J., & Zhang, Y. (2006). A dynamic model of interaction between reliance on automation and cooperation in multi-operator multi-automation situations. *International Journal of Industrial Ergonomics*, *36*(5), 511–526. doi:10.1016/j.ergon.2006.01.013

Gawande, A. (2009). *The Checklist Manifesto*. New York, NY: Henry Holt and Company, LLC.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*(2), 169–79. doi:10.1006/jtbi.2000.2111

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings.* (pp. 55–60). Millbrae, CA: Ieee. doi:10.1109/ROMAN.2003.1251796

Goldstein, J., Pevehouse, J., Gerner, D., & Telhami, S. (2001). Reciprocity, Triangularity, and Cooperation in the Middle East, 1979-97. *Journal of Conflict Resolution*, *45*(October), 594–620. doi:10.1177/0022002701045005003

Gonzalez, C., Vanyukov, P., & Martin, M. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior*, *21*(2), 273–286. doi:10.1016/j.chb.2004.02.014

Gorman, J. (2011). Team coordination dynamics and the interactive approach: Emerging evidence and future work. In *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems* (pp. 298–307). Springer Berlin Heidelberg.

Gorman, J., Cooke, N., & Amazeen, P. (2010). Training Adaptive Teams. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(2), 295–307. doi:10.1177/0018720810371689

Gouldner, A. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, *25*(2), 161–178. doi:10.2307/2092623

Grant, R. (1996). Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organization Science*, *7*(4), 375–387.

Groom, V., & Nass, C. (2007). Can robots be teammates? *Interaction Studies*, *8*(3), 285–301. doi:10.1075/gest.8.3.02str

Grundgeiger, T., & Sanderson, P. (2009). Interruptions in healthcare: Theoretical views. *International Journal of Medical Informatics*, *78*(5), 293–307.

doi:10.1016/j.ijmedinf.2008.10.001

Hancock, P., Billings, D., Schaefer, K., Chen, J., de Visser, E., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517–527. doi:10.1177/0018720811417254

Ho, G., Wheatley, D., & Scialfa, C. (2005). Age differences in trust and reliance of a medication management system. *Interacting with Computers*, *17*(6), 690–710. doi:10.1016/j.intcom.2005.09.007

Hoc, J.-M. (2000). From human-machine interaction to human-machine cooperation. *Ergonomics*, *43*(7), 833–843.

Hoc, J.-M. (2001). Towards a cognitive approach to human–machine cooperation in dynamic situations. *International Journal of Human-Computer Studies*, *54*, 509–540. doi:10.1006/ijhc.2000.0454

Hoc, J.-M. (2013). Human-Machine Cooperation. In J. D. Lee & A. Kirlik (Eds.), *The Oxford Handbook of Cognitive Engineering* (pp. 395–403). Oxford: Oxford University Press.

Hoff, K., & Bashir, M. (2014). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. doi:10.1177/0018720814547570

Hoffman, G., & Ju, W. (2014). Designing robots with movement in mind. *Journal of Human-Robot Interaction*, *3*(1), 89. doi:10.5898/JHRI.3.1.Hoffman

Hollnagel, E. (2012). A Tale of Two Safeties. *Nuclear Safety and Simulation*, *4*(1), 1–9.

Hollnagel, E., Woods, D., & Leveson, N. (Eds.). (2006). *Resilience Engineering: Concepts and Precepts*. Hampshire, England: Ashgate Publishing Limited.

Horvitz, E. (1999). Principles of Mixed-Imitative User Interfaces. *Conference on Human Factors in Computing Systems*, (May), 159–166.

Janssen, M. (2008). Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents. *Journal of Economic Behavior & Organization*, *65*(3-4), 458–471. doi:10.1016/j.jebo.2006.02.004

Janssen, O., & Veenstra, C. (1999). How task and person conflict shape the role of positive interdependence in management teams. *Journal of Management*, *25*(2), 117–141. doi:10.1016/S0149-2063(99)80006-3

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, *47*(2), 263–291.

Kaplan, F. (2001). Artificial attachment: Will a robot ever pass ainsworth s strange situation test. *Proceedings of Humanoids*, 125–132.

Kaplan, M. (2013). The steely, headless king of Texas Hold 'Em. *The New York Times Magazine*.

Kazi, T., Stanton, N., Walker, G., & Young, M. (2007). Designer driving: Drivers' conceptual models and level of trust in adaptive cruise control. *International Journal of Vehicle Esign*, *45*(August 2015), 339–360. doi:10.1504/IJVD.2007.014909

Kelley, H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107–

128. doi:10.1037/h0034225

Kelley, H., Holmes, J., Kerr, N., Reis, H., Rusbult, C., & Van Lange, P. (2003). *An Atlas of Interpersonal Situations*. Cambridge, UK: Cambridge University Press.

Kim, D., Ferrin, D., & Rao, H. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems*, (44), 544–564.

Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: Why an "aid" can (and should) go unused. *Human Factors*, *35*(2), 221–242. doi:10.1177/001872089303500203

Knight, W. (2013). Smart robots can now work right next to auto workers. *MIT Technology Review*. Retrieved from http://www.technologyreview.com/news/518661/smart-robots-can-now-work-right-next-to-auto-workers/

Kramer, R. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*, 569–598. doi:10.1146/annurev.psych.50.1.569

Lasota, P., & Shah, J. (2015). Analyzing the effects of human-aware motion planning on close-proximity human-robot collaboration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(1), 21–33. doi:10.1177/0018720814565188

Lee, J., Knox, B., & Breazeal, C. (2011). Modeling the dynamics of nonverbal behavior on interpersonal trust for human-robot interactions.

Lee, J., & See, K. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80.

Levy, S. (2012). Can an algorithm write a better news story than a human reporter? *Wired*.

Lewicki, R., McAllister, D., & Bies, R. (1998). Trust and distrust: New relationships and realities. *The Academy of Management Review*, *23*(3), 438. doi:10.2307/259288

Li, X., Hess, T., & Valacich, J. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems*, *17*(1), 39–71. doi:10.1016/j.jsis.2008.01.001

Lyons, J., & Stokes, C. (2011). Human-Human Reliance in the Context of Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(1), 112–121. doi:10.1177/0018720811427034

Madhavan, P., & Wiegmann, D. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. doi:10.1080/14639220500337708

Madhavan, P., Wiegmann, D., & Lacson, F. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. In *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics SOciety*. Santa Monica, CA.

Main, B., Taubitz, M., & Wood, W. (2008). You cannot get lean without safety: Understanding the common goals. *Professional Safety*, *53*(1), 38–42.

Malone, T., & Crowston, K. (1994). The interdisciplinary study of coordination. *ACM Computing Surveys*, *26*(1), 87–119. doi:10.1145/174666.174668

Markussen, T. (2009). Bloody robots as emotional design: How emotional structures may change expectations of technology use in hospitals. *International Journal of Design*, *3*(2), 27–39.

Martin, J., Gonzalez, C., Juvina, I., & Lebiere, C. (2013). A description-experience gap in social interactions: Information about interdependence and its effects on cooperation. *Journal of Behavioral Decision Making*, *362*(December 2013), 349–362. doi:10.1002/bdm.1810

Mayer, R., Davis, J., & Schoorman, F. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11832–11835. doi:10.1073/pnas.211415698

McFarlane, D. (1999). Coordinating the interruption of people in human-computer interaction. In *30th International Conference on Human-Computer Interaction* (pp. 295–303). Edinburgh, UK: IOS Press.

McFarlane, D., & Latorella, K. (2002). The Scope and Importance of Human Interruption in Human-Computer Interaction Design. *Human-Computer Interaction*, *17*(1), 1–61. doi:10.1207/S15327051HCI1701_1

Mcknight, D., Carter, M., Thatcher, J., & Clay, P. (2011). Trust in a specific technology: An investigation in its components and measures. *ACM Transactions on …*, *2*(2), 1–12. doi:10.1145/1985347.1985353

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*(2), 196–204.

Meyer, J., & Lee, J. (2013). Trust, reliance, and compliance. In J. D. Lee & A. Kirlik (Eds.), *The Oxford Handbook of Cognitive Engineering* (pp. 109–124). Oxford: Oxford University Press.

Miller, C., Funk, H., Goldman, R., Meisner, J., & Wu, P. (2005). Implications of adaptive vs. adaptable UIs on decision making: Why "automated adaptiveness" is not always the right answer. *Proceedings of the 1st International Conference on Augmented Cognition*, 1180–1189. doi:10.1.1.148.6963

Molm, L., Schaefer, D., & Collett, J. (2007). The value of reciprocity. *Social Psychology Quarterly*, *70*(2), 199–217.

Molm, L., Takahashi, N., & Peterson, G. (2000). Risk and trust in social exchange: An experimental test of a classical proposition. *American Journal of Sociology*, *105*(5), 1396–1427.

Montague, E., Kleiner, B., & Winchester III., W. (2009). Empirically understanding trust in medical technology. *International Journal of Industrial Ergonomics*, *39*(4), 628–634. doi:10.1016/j.ergon.2009.01.004

Muir, B. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5-6), 527–539. doi:10.1016/S0020-7373(87)80013-5

Mutlu, B., & Forlizzi, J. (2008). Robots in organizations: The role of workflow , social , and

environmental factors in human-robot interaction. *Human-Computer Interaction Institute, Paper 36*.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., & Ishiguro, H. (2012). Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems*, *1*(2), 1–33. doi:10.1145/2070719.2070725

Nass, C., & Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, MA: MIT Press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. doi:10.1111/0022-4537.00153

Nass, C., Steuer, J., Tauber, E., & Reeder, H. (1993). Anthropomorphism, agency, and ethopoeia: Computers as social actors. *Computer-Human Interaction (CHI) Conference 1993*, 111–112. doi:http://doi.acm.org/10.1145/259964.260137

Nemeth, C., Wears, R., Woods, D., Hollnagel, E., & Cook, R. (2008). Minding the gaps: Creating resilience in health care. In *Advances in Patient Safety: New Directions and Alternative* (Vol. 3). Rockville, MD: Agency for Healthcare Research and Quality. doi:NBK43670 [bookaccession]

Nowak, M., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(October), 1291–11298. doi:Doi 10.1038/31225

Omodei, M., & Wearing, A. (1995). The Fire-Chief Microworld Generating Program - an Illustration of Computer-Simulated Microworlds as an Experimental Paradigm for Studying Complex Decision-Making Behavior. *Behavior Research Methods, Instruments, & Computers*, *27*(3), 303–316.

Ostrom, E. (1999). Revisiting the commons: Local lessons, global challenges. *Science*, *284*(5412), 278–282. doi:10.1126/science.284.5412.278

Ostrom, E. (2000). Collective action and the evolution of social norms. *The Journal of Economic Perspectives*, *14*(3), 137–158.

Parasuraman, R., & Miller, C. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55.

Parasuraman, R., Molloy, R., & Singh, I. (1993). Performance Consequences of Automation-Induced "Complacency." *The International Journal of Aviation Psychology*, *3*(1), 1–23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. doi:10.1518/001872097778543886

Parasuraman, R., Sheridan, T., & Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–97. doi:10.1109/3468.844354

Parks, C., Henager, R., & Scamahorn, S. (1996). Trust and reactions to messages of intent in social dilemmas. *Journal of Conflict Resolution*, *40*(1), 134–151.

Porges, S. (2001). The polyvagal theory: Phylogenetic substrates of a social nervous system. *International Journal of Psychophysiology*, *42*(2), 123–146. doi:10.1016/S0167-8760(01)00162-3

Porges, S. (2007). The polyvagal perspective. *Biological Psychology*, *74*(2), 116–143.

doi:10.1016/j.str.2010.08.012.Structure

Pritchett, A., Kim, S., & Feigh, K. (2013). Measuring Human-Automation Function Allocation. *Journal of Cognitive Engineering and Decision Making*, *8*(1), 52–77. doi:10.1177/1555343413490166

R Core Team. (2014). A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rankin, A., Lundberg, J., Woltjer, R., Rollenhagen, C., & Hollnagel, E. (2013). Resilience in everyday operations: A framework for analyzing adaptations in high-risk work. *Journal of Cognitive Engineering and Decision Making*, *8*(1), 78–97. doi:10.1177/1555343413498753

Riedl, R., Mohr, P., Kenning, P., Davis, F., & Heerkeren, H. (2011). Trusting humans and avatars: Behavioral and neural evidence. In *Thirty Second International Conference on Information Systems*. Shanghai, China.

Rivera-Rodriguez, A., & Karsh, B.-T. (2010). Interruptions and distractions in healthcare: review and reappraisal. *Quality & Safety in Health Care*, *19*(4), 304–12. doi:10.1136/qshc.2009.033282

Robert, L., Dennis, A., & Hung, Y. (2009). Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *Journal of Management Information Systems*, *26*(2), 241–279. doi:10.2753/MIS0742-1222260210

Robinette, P., Wagner, A., & Howard, A. (2013). Building and maintaining trust between humans and guidance robots in an emergency. In *AAAI Spring Symposium, Stanford University* (pp. 78–83).

Rochlin, G., La Porte, T., & Roberts, K. (1987). The self-designing high-reliability organization: Aircraft carrier flight operations at sea. *Naval War College Review*, *40*(4), 76–90.

Roloff, M. (1981). *Interpersonal Communication: The Social Exchange Approach*. Beverly Hills, CA: SAGE Publications, Inc.

Rotter, J. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651–65. doi:10.1111/j.1467-6494.1967.tb01454.x

Rotter, J. (1971). Generalized expectancies for interpersonal trust. *American Psychologist*, *26*(5), 443–452.

Rusbult, C., & Van Lange, P. (2003). Interdependence, interaction, and relationships. *Annual Review of Psychology*, *54*, 351–375. doi:10.1146/annurev.psych.54.101601.145059

Salas, E., Burke, C., & Janis, A. (2000). Teamwork: Emerging principles. *International Journal of Management Reviews*, *2*(4), 339–356.

Salas, E., Cooke, N., & Rosen, M. (2008). On teams, teamwork, and team performance: Discoveries and developments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 540–547. doi:10.1518/001872008X288457.

Salas, E., Wilson, K., Murphy, C., King, H., & Salisbury, M. (2008). Communicating, coordinating, and cooperating when lives depend on it: Tips for teamwork. *Joint Commission Journal on Quality and Patient Safety / Joint Commission Resources*, *34*(6), 333–41.

Sandoval, E., Brandstetter, J., Obaid, M., & Bartneck, C. (2016). Reciprocity in human robot interaction: A quantitative approach through The prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics, 8,* 303–317. doi:10.1007/s12369-015-0323-x

Schelling, T. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schelling, T. (1973). Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *Journal of Conflict Resolution, 17*(3), 381–428. doi:10.1177/002200277301700302

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *The Behavioral and Brain Sciences, 36,* 393–414. doi:10.1017/S0140525X12000660

Sheridan, T. (2008). Risk, human error, and system resilience: fundamental ideas. *Human Factors, 50*(3), 418–426. doi:10.1518/001872008X250773

Sheridan, T., & Verplank, W. (1978). *Human and computer control of undersea teleoperators*. Arlington, VA.

Simon, H. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118. doi:10.2307/1884852

Smith, M., Carayon, P., Sanders, K., Lim, S., & Legrande, D. (1992). Employee stress and health complaints in jobs with and without electronic performance monitoring. *Applied Ergonomics, 23*(1), 17–27.

Staples, D., & Webster, J. (2008). Exploring the effects of trust, task interdependence and virtualness on knowledge sharing in teams. *Information Systems Journal*, (18), 617–640.

Stephens, R., Woods, D., Branlat, M., & Wears, R. (2011). Colliding dilemmas: Interactions of locally adaptive strategies in a hospital setting. In *Fourth Symposium on Resilience Engineering*. Sophia Antipolis, France.

Takayama, L. (2009). Making sense of agentic objects and teleoperation: In-the-moment and reflective perspectives. *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*, 239–240. doi:10.1145/1514095.1514155

Takayama, L., Dooley, D., & Ju, W. (2011). Expressing thought: Improving robot readability with animation principles. In *Proceedings of Human-Robot Interaction Conference* (pp. 69–76). Lausanne, CH. doi:10.1145/1957656.1957674

Takayama, L., Groom, V., & Nass, C. (2009). I'm sorry, Dave: I'm afraid I won't do that: Social aspects of human-agent conflict. In *CHI 2009 - Studying Intelligent Systems* (pp. 2099–2107). Boston, MA: ACM. doi:10.1145/1518701.1519021

Terada, K., Shamoto, T., Mei, H., & Ito, A. (2007). Reactive movements of non-humanoid robots cause intention attribution in humans. *IEEE International Conference on Intelligent Robots and Systems*, 3715–3720. doi:10.1109/IROS.2007.4399429

Thibaut, J., & Kelley, H. (1959). *The Social Psychology of Groups*. New York: John Wiley & Sons, Inc.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Van Wezel, W., Cegarra, J., & Hoc, J.-M. (2011). Allocating functions to human and

algorithm in scheduling. In J. C. Fransoo, T. Wafler, & J. Wilson (Eds.), *Behavioral Operations in Planning and Scheduling* (pp. 339–370). Heidelberg: Springer.

Verberne, F., Ham, J., & Midden, C. (2012). Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(5), 799–810. doi:10.1177/0018720812443825

Verberne, F., Ham, J., & Midden, C. (2015). Trusting a virtual driver that looks, acts, and thinks like you. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(5), 895–909. doi:10.1177/0018720815580749

Wagner, A. (2013). Developing Robots that Recognize When They Are Being Trusted. In *AAAI Spring Symposium* (pp. 84–89). Stanford University.

Wagner, A., & Arkin, R. (2008). Analyzing social situations for human–robot interaction. *Interaction Studies*, *9*(2), 277–300. doi:10.1075/is.9.2.07wag

Wagner, A., & Arkin, R. (2011). Recognizing situations that demand trust. In *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication* (pp. 7–14). Atlanta. doi:10.1109/ROMAN.2011.6005228

Walji, M., Brixey, J., Johnson-Throop, K., & Zhang, J. (2004). A theoretical framework to understand and engineer persuasive interruptions. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, *66*(5), 297–333. doi:10.1037/h0040934

Wickham, H. (2009). Ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag.

Wickham, H., & Francois, R. (2015). dplyr: A Grammar of Data Manipulation. *R Package Version 0.4.3*.

Wicks, A., Berman, S., & Jones, T. (1999). The structure of optimal trust: Moral and strategic implications. *Academy of Management Review*, *24*(1), 99–116.

Williams, K. (2010). Dyads can be groups (and often are). *Small Group Research*, *41*(2), 268–274. doi:10.1177/1046496409358619

Woods, D. (2004). Conflicts between learning and accountability in patient safety. *DePaul Law Review*, *54*, 485–502.

Woods, D. (2015). Four concepts for resilience and the implications for the future of resilience engineering. *Reliability Engineering & System Safety*, *141*, 5–9. doi:10.1016/j.ress.2015.03.018

Woods, D., & Branlat, M. (2010). Basic patterns in how adaptive systems fail. In E. Hollnagel, J. Pariès, D. D. Woods, & J. Wreathall (Eds.), *Resilience Engineering in Practice: A Guidebook* (pp. 127–144). Farnham, UK, UK: Ashgate.

Woods, D., & Hollnagel, E. (2005). *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*. Boca Raton, FL: CRC Press.

Woods, D., & Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. Boca Raton, FL: CRC Press.

Woods, D., Johannesen, L., Cook, R., & Sarter, N. (1994). *Behind human error: Cognitive systems, computers and hindsight*. (A. Schopper, Ed.)*CSERIAC SOAR 94-01*. Wright-Patterson Air Force Base, OH.

Woods, D., & Patterson, E. (2001). How unexpected events produce an escalation of cognitive and coordinative demands. In P. Hancock & P. Desmond (Eds.), *Stress Workload and Fatigue* (pp. 290–304). Hillsdale, NJ: Lawrence Erlbaum Associates.

Woods, D., Tittle, J., Feil, M., & Roesler, A. (2004). Envisioning human–robot coordination in future operations. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 34*(2), 210–218. doi:10.1109/TSMCC.2004.826272

Young, P. (1996). The economics of convention. *The Journal of Economic Perspectives*, *10*(2), 105–122.

Zhang, S., & Adam, M. (2012). Humans versus agents: Competition in financial markets of the 21st century. In *Thirty Third International Conference on Information Systems*. Orlando.

Zheng, J., Veinott, E., Bos, N., Olson, J., & Olson, G. (2002). Trust without touch: jumpstarting long-distance trust with initial social activities. *CHI Letters*, *4*(1), 141–146.

Zieba, S., Polet, P., Vanderhaegen, F., & Debernard, S. (2009). Resilience of a human-robot system using adjustable autonomy and human-robot collaborative control. *International Journal of Adaptive and Innovative Systems, 1*(1), 13–29. doi:10.1504/IJAIS.2009.022000

Zieba, S., Polet, P., Vanderhaegen, F., & Debernard, S. (2010). Principles of adjustable autonomy: A framework for resilient human-machine cooperation. *Cognition, Technology and Work, 12*(3), 193–203. doi:10.1007/s10111-009-0134-7

Appendices

Appendix A: Propensity to Trust Questionnaire

The questionnaire as follows was administered at the start of the study. Each item was accompanied by a six-point scale with the options: “Strongly inaccurate”, “Inaccurate”, “Somewhat inaccurate”, “Somewhat accurate”, “Accurate”, and “Strongly accurate”. Negatively scored questions are marked with a (-), which were not shown on the actual questionnaire, which was administered on Qualtrics, a web-based platform.

Instructions: Please rate the extent that each item describes you.

1. I listen to my conscience.
2. I anticipate the needs of others.
3. I respect others.
4. I can get along with most people.
5. I have always been completely fair to others.
6. I stick to the rules.
7. I believe that laws should be strictly enforced.
8. I have a good word for everyone.
9. I value cooperation over competition.
10. I return extra change when a cashier makes a mistake.
11. I would never cheat on my taxes.
12. I follow through with my plans.
13. I believe that people are basically moral.
14. I finish what I start.
15. I retreat from others. (-)
16. I am filled with doubts about things. (-)
17. I feel short-changed in life. (-)
18. I avoid contact with others. (-)
19. I believe that most people would lie to get ahead. (-)
20. I find it hard to forgive others. (-)

Appendix B: Task Interdependence Questionnaire

Task interdependence is measured through five items from Staples and Webster's (2008, p. 640) six-item scale for task interdependence, based on work by Bishop and Scott (2000), and Janssen (1999). An item on team communication was removed because team communication was severely limited in this study. Other adaptations to the items replace "team members" with more related terms for this study, such as "neighboring hospital agent". The questionnaire as follows was administered following the microworld trials. Because the microworld task is highly interdependent, and given explicit training on the task and task goals, it was generally expected that participants would respond accordingly. Each item was rated on a six- point scale with the options: "Strongly disagree", "Disagree", "Somewhat disagree", "Somewhat agree", "Agree", "Strongly agree". The following questionnaire was administered on Qualtrics, a web-based survey and questionnaire platform.

Instructions: Please select the answer that best describes how you feel about the task you just completed, including interacting with the neighboring hospital.

1. I frequently must coordinate my efforts with the neighboring hospital.
2. Goal attainment for one hospital helped goal attainment for the other hospital.
3. To achieve high performance, it was important to rely on each other.
4. The tasks performed by the different hospital schedulers were related to one another.
5. Success for one hospital implied success for the other hospital.

Appendix C: Demographic Questionnaire

To assess the sample population, a demographics questionnaire was administered at the end of the study. The questions assessed age, gender, education level and profession, experience with computers, and experience with video games to control for these potential factors in the results. The questionnaire was roughly formatted as follows, and administered on Qualtrics, a web-based survey and questionnaire platform.

Age: _____

Gender (select): Male / Female / Other

If a current student, please write in the following:

College _____
Major _____
Degree in pursuit of_____
Number of years pursuing this degree _____

If employed, please write in your occupation: _____

Highest level of education (select one):
____ Some high school or less
____ High school diploma
____ 2-year college degree or trade school
____ 4-year college degree
____ Masters degree
____ Professional degree
____ Doctorate degree

I use a computer (select one):
____ Daily
____ Every couple days
____ Once a week
____ Every couple weeks
____ Less than once a month
____ Never

I use the computer for (check all that apply):
____ Internet searching
____ Email
____ Document processing
____ Computer Games
____ Other (Please specify _____)

I play the following categories of video games (check all that apply):
___ Sports
___ Real-time strategy
___ First person shooter
___ Racing
___ Role-playing (RPG)
___ Flight simulator
___ Other (Please specify _____)

I use the following technologies (check all that apply):
____ Basic mobile phone
____ Mobile smart phone/PDA/Blackberry/MP3 player
____ Touch screens (on commercial tablets, phones, or GPS navigation devices)

How much experience do you have playing video games?
____ None
____ Less than 1 year
____ 1 – 2 years
____ More than 2 years

I play video games (select one):
____ Daily
____ A few times a week
____ A few times a month
____ A few times a year
____ Less than a few times a year
____ Never

Appendix D: Baseline Tempo

It was opted not to test a baseline tempo pattern, due to the suspicion that an appropriate baseline condition may not exist. A dynamic, random tempo pattern would allow testing of cooperation across variation in tempo patterns, but trust in automation research suggests that higher variability in the environment, thus increased uncertainty, may lead to stronger reliance on automation. A random tempo pattern might indeed lead to higher perceived variability in the environment, and a separate study focusing on people's perceptions of environmental uncertainty in the different conditions would be needed before considering a dynamic random tempo as a viable baseline. A static tempo pattern would not be an appropriate baseline either because a static tempo would not represent a dynamic task environment by definition, and the predictability of a static tempo may lead to boredom or disengagement in the task. Furthermore, a static tempo would potentially direct focus to individual differences between participants, which is not the purpose of this study. The differences between people with higher and lower skills, e.g. due to speed or strategy, may be emphasized in a static tempo compared to a dynamic tempo.

Appendix E: Two Outliers

Two participants were removed as outliers, due to closer inspection of their data which showed neither gave any staff to the agent during their experimental trials. These data were not included in the analysis because such behavior would greatly skew the results – mainly, there was no interaction data on the participants' side to inspect, and receiving staff from the agent without ever returning any would lead to poor joint scores due to the microworld design. Additional participants were collected at the end of the study period to reach the target sample size of $N = 50$.

It was uncertain why the two participants did not give staff; both scored relatively high on the disposition to trust scale (at least four out of six, where six indicates more trusting) and high on the task interdependence scale (at least five out of six, where six indicates they perceived the task as very interdependent). The training trial each participant experienced involved the researcher observing that every participant was exposed to every potential function of the interface, so that understanding the user interface would not be an issue during experimental trials. This included the researcher suggesting to participants they try each action and then observe them doing so correctly, if some participants did not attempt certain actions – like giving staff – on their own during the 2 minute practice trial. In addition, in researcher notes taken during post-experiment debriefing, both participants mentioned they believe they gave less staff than the agent and were less cooperative, indicating they accurately perceived their comparative performance, and understood the joint goal and mechanisms of the microworld. One participant mentioned, "I was going so fast I forgot the other agent" and the other, "It was a busy hospital, should help it more with my medical staff"; however, it should be noted that neither of these responses were unique to these two participants.