The Fundamental Principle of Coactive Design: Interdependence Must Shape Autonomy

Matthew Johnson[1,2], Jeffrey M. Bradshaw[1], Paul J. Feltovich[1],
Catholijn M. Jonker[2], Birna van Riemsdijk[2], Maarten Sierhuis[2,3]

[1] Florida Institute for Human and Machine Cognition (IHMC), Pensacola, Florida, USA
[2]EEMCS, Delft University of Technology, Delft, The Netherlands
[3]PARC, Palo Alto, California, USA
{mjohnson, jbradshaw, pfeltovich}@ihmc.us
{c.m.jonker and m.b.vanriemsdijk}@tudelft.nl
maarten.sierhuis@parc.com

Abstract. This article presents the fundamental principle of *Coactive Design*, a new approach being developed to address the increasingly sophisticated roles for both people and agents in mixed human-agent systems. The fundamental principle of Coactive Design is that the underlying *interdependence* of participants in joint activity is a critical factor in the design of human-agent systems. In order to enable appropriate interaction, an understanding of the potential interdependencies among groups of humans and agents working together in a given situation should be used to shape the way agent architectures and individual agent capabilities for autonomy are designed. Increased effectiveness in human-agent teamwork hinges not merely on trying to make agents more independent through their autonomy, but also in striving to make them more capable of sophisticated *interdependent* joint activity with people.

Keywords: Coactive, autonomy, interdependence, joint activity

1 Introduction

Researchers and developers continue to pursue increasingly sophisticated roles for agents.[1] Envisioned roles include caretakers for the homebound, physician assistants, coworkers and aides in factories and offices, and servants in our homes. Not only are the agents themselves increasing in their capabilities, but also the composition of human-robot systems is growing in scale and heterogeneity. All these requirements showcase the importance of robots transitioning from today's common modes of reliance, where they are frequently operated as mere teleoperated tools, to more sophisticated partners or teammates [1, 2].

Direct teleoperation and complete autonomy are often thought of as two extremes on a spectrum. Researchers in human-agent interaction have typically seen themselves as investigating the middle ground between these extremes. Such research has gone under various names, including mixed-initiative interaction [3], adjustable autonomy

[1] Throughout the article we will use the terms "agent" and "robot" interchangeably to mean any artificial actor.

[4], collaborative control [5], and sliding autonomy [6]. Each of these approaches attempts to keep the human-agent system operating at a “sweet spot” between the two extremes. As the names of these approaches suggest, researchers understand that the ideal is not a fixed location along this spectrum but may need to vary dynamically along the spectrum as context and resources change. Of importance to our discussion is the fact that these approaches, along with traditional planning technologies at the foundation of intelligent systems, typically take an autonomy-centered perspective, focusing mainly on the problems of control and task allocation when agents and humans attempt to work together.

In contrast to these autonomy-centered approaches, Coactive Design is a teamwork-centered approach. The concept of teamwork-centered autonomy was addressed by Bradshaw *et al.* [7]. It takes as a beginning premise that joint activity of a consequential nature often requires people to work in close and continuous interaction with autonomous systems, and hence adopts the stance that the processes of understanding, problem solving and task execution are necessarily incremental, subject to negotiation, and forever tentative.

The overall objective of our work in Coactive Design is to describe and, insofar as possible, empirically validate design principles and guidelines to support joint activity in human-agent systems. Though these principles and guidelines are still under development, our research has progressed to the point where we are ready to present the fundamental principle that serves as the foundation for our approach. The fundamental principle of Coactive Design recognizes that the underlying *interdependence* of participants in joint activity is a critical factor in the design of human-agent systems. In order to enable appropriate interaction, an understanding of the potential interdependencies among groups of humans and agents working together in a given situation should be used to shape the way agent architectures and individual agent capabilities for autonomy are designed. We no longer look at the primary problem of the research community as simply trying to make agents more independent through their autonomy. Rather, in addition, we strive to make them more capable of sophisticated interdependent joint activity with people.

This article will begin by an overview of different usages of the term *autonomy* in the agent and robot literature. We provide a rationale for our belief that a new approach to human-agent system design is needed in the context of prior research and its associated challenges. Next we introduce some of the concepts important to the Coactive Design approach and present different aspects of its fundamental principle. We discuss relevant experimental work to date that has begun to demonstrate our claims. Finally, we close with a summary of the work.

2 Defining Autonomy

Autonomy has two basic senses in everyday usage. The first sense, self-sufficiency, is about the degree to which an entity is able to take care of itself. Bradshaw [8] refers to this as the *descriptive dimension* of autonomy. Similarly, Castelfranchi [9] referred to this as one of the two aspects of *social autonomy* that he called *independence*. People usually consider robot autonomy in this sense in relation to a particular task. For example, a robot may be able to navigate autonomously, but only in an office environment. The second sense refers to the quality of self-directedness, or the degree of freedom from outside constraints (whether social or environmental), which Bradshaw calls the *prescriptive dimension* of autonomy. Castelfranchi referred to this as autonomy of delegation and considered it another form of *social autonomy*. For robots, this usually means freedom from human input or intervention during a particular task.

In the following section, we will describe some of the more prominent approaches to improve human-robot system effectiveness.[2] To avoid the ambiguity often found in the agent literature, we will use the terms *self-sufficiency* and *self-directedness* in our discussion.

3 Prior Work

3.1 Function Allocation and Supervisory Control

The concept of automation—which began with the straightforward objective of replacing whenever feasible any task currently performed by a human with a machine that could do the same task better, faster, or cheaper—became one of the first issues to attract the notice of early human factors researchers. These researchers attempted to systematically characterize the general strengths and weaknesses of humans and machines [10]. The resulting discipline of *Function Allocation* aimed to provide a rational means of determining which system-level functions should be carried out by humans and which by machines. Sheridan proposed the concept of *Supervisory Control* [11], in which a human oversees one or more autonomous systems, statically allocating tasks to them. Once control is given to the system, it is ideally expected to complete the tasks without human intervention. The designer's job is to determine what needs to be done and then provide the agent the capability (i.e., self-sufficiency) to do it. Therefore, this approach to achieving autonomy is shaped by a system's self-sufficiency.

3.2 Adaptive, Sliding, or Adjustable Autonomy

Over time it became plain to researchers that things were not as simple as they first appeared. For example, the suitability of a particular human or machine to take on a particular task may vary by time and over different situations; hence the need for methods of function allocation that are dynamic and adaptive. Dorais [12] defines "adjustable autonomy" as "the ability of autonomous systems to operate with dynamically varying levels of independence, intelligence and control." Dias [13] uses a similar definition for the term "sliding autonomy." Sheridan discusses "adaptive automation," in which the system must decide at runtime which functions to automate and to what extent. We will use the term *adjustable autonomy* as a catch-all to refer to this concept, namely, a change in agent autonomy—in this case the self-directedness aspect—to some appropriate level, based on the situation. The action of adjustment may be initiated by the human, by the agent framework, or by the agent itself.

[2] Parts of our discussion of this topic are adapted from [8].

It is evident that such approaches are autonomy-centered, with the focus being on task assignment, control, and level of independence. Autonomy, in this case, is shaped exclusively by varying levels of self-directedness. One very important concept emphasized by these approaches is adaptivity, a quality that will be important in the operation of increasingly-sophisticated intelligent systems.

3.3 Mixed-Initiative Interaction

Mixed-initiative approaches evolved from a different research community, but share some similar ideas and assumptions. Allen defines mixed-initiative as "a flexible interaction strategy, where each agent can contribute to the task what it does best" [3]. In Allen's work, the system is able to reason about which party should initiate action with respect to a given task or communicative exchange. In a similar vein, Myers and Morley describe a framework called "Taskable Reactive Agent Communities (TRAC) [14] that supports the directability of a team of agents by a human supervisor by modifying task guidance." Directability or task allocation is once again the central feature of the approach. Murphy [15] also uses the term "mixed-initiative" to describe their attention-directing system, the goal of which is to get the human to assume responsibility for a task when a robot fails.

Mixed-initiative interaction is also essentially autonomy-centered. Its usual focus is on task assignment or the authority to act and, as such, varying self-directedness is used to shape the operation of the autonomous system. Mixed-initiative interaction contributes the valuable insight that joint activity is about interaction and negotiation, and that dynamic shifts in control may be useful.

3.4 Collaborative Control

Collaborative Control is an approach proposed by Fong [5] that uses human-robot dialogue (i.e., queries from the robot and the subsequent presence or absence of a responses from the human), as the mechanism for adaptation. As Fong states, "Collaborative control... allows robots to benefit from human assistance during perception and cognition, and not just planning and command generation" [5]. Collaborative Control is a first step toward Coactive Design, introducing the idea that both parties may participate simultaneously in the same action. Here the ongoing interdependence of the human and the robot in carrying out a navigation task is used to shape the design of autonomous capabilities. The robot was designed to enable the human to provide assistance in the perceptual and cognitive parts of the task. The robotic assistance is not strictly required, so we are not merely talking about self-sufficiency. The key point is that the robotic assistance in this case is an integral part of the robot design and operation. We have adopted and extended some of the ideas from Collaborative Control as we have developed the Coactive Design approach.

3.5 How Autonomy Has Been Characterized in Former Research

One way to gain insight into the predominant perspectives in a research community is to review how it categorizes and describes its own work. This provides a test of our claim that prior work in agents and robots has been largely autonomy-centered.

The general drift is perhaps most clearly seen in the work of researchers who have tried to describe different "levels" of autonomy. For example, Yanco [16] characterized autonomy in terms of the amount of intervention required. For example, full teleoperation is 100% intervention and 0% automation. On the other hand, tour guide robots are labelled 100% autonomous and 0% intervention. The assumption in this model is that intervention only occurs when the robot lacks self-sufficiency. However, identifying the "percentage" of intervention is a very subjective matter except when one is at the extreme ends of the spectrum. Similarly Parasuraman and Sheridan [17] provide a list of levels of autonomy shown in figure 1.

HIGH	10. The computer decides everything, acts autonomously, ignoring the human.
	9. informs the human only if it, the computer, decides to
	8. informs the human only if asked, or
	7. executes automatically, then necessarily informs the human, and
	6. allows the human a restricted time to veto before automatic execution, or
	5. executes that suggestion if the human approves, or
	4. suggests on alternative
	3. narrows the selection down to a few, or
	2. The computer offers a complete set of decision/action alternatives, or
LOW	1. The computer offers no assistance: human must take all decisions and actions

Fig. 1. Levels of Automation [17].

Sheridan's scale is clearly autonomy-centered, as noted by Goodrich and Schultz [18]. Specifically it focused on the self-directedness aspect of autonomy. In response to the limitations of Sheridan's scale, Goodrich and Schultz [18] developed a scale that attempts to focus on levels of interaction rather than of automation (figure 2).

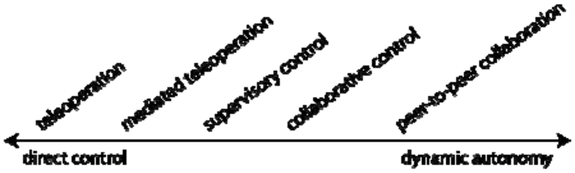


Fig. 2. Levels of autonomy with an emphasis on human interaction [18].

Though Goodrich and Schultz rightfully recognized that something more than the previous autonomy-centered characterizations of the field needed to be captured, in reality the left-to-right progress of the scale provides little more than a historical summary of robot research, with peer-to-peer collaboration as the next step. The label of the right end of the spectrum, "dynamic autonomy," reveals that this scale is, like the others discussed previously, autonomy-centered.

Bradshaw has characterized autonomy in terms of multiple dimensions rather than a single one-dimensional scale of levels [8]. The descriptive and prescriptive aspects of autonomy discussed above capture two of these primary dimensions. He also argues that the measurement of these dimensions should be specific to task and situation, since an agent may be self-directed or self-sufficient in one particular task or situation, but not in another.

Castelfranchi suggested dependence as the complement of autonomy [9] and attempts to capture several dimensions of autonomy in terms of the autonomy vs. dependence of various capabilities in a standard Procedural Reasoning System (PRS) architecture. These include information, interpretation, know-how, planning, plan discretion, goal dynamics, goal discretion, motivation, reasoning, monitoring, and skill autonomy. Like Bradshaw, Castelfranchi recognizes that autonomy is not a monolithic property, but should be measured with respect to different aspects of the agent. Castelfranchi put it this way: "any needed resource or power within the action-perception loop of an agent defines a possible dimension of dependence or autonomy" [9].

3.6 Challenges of Autonomy-Centered Approaches

We now describe the most common challenges faced by autonomy-centered approaches in the context of both senses of autonomy. Since the capability to perform a task and the authority to perform a task are orthogonal concepts, we separate these two dimensions onto separate axes, as in figure 3. Together these two axes represent an autonomy-centered plane of robotic capabilities. The *self-sufficiency* axis represents the degree to which a robot can perform a task by itself. "Low" indicates that the robot is not capable of performing the task without significant help. "High" indicates that the robot can perform the task reliably without assistance. The *self-directedness* axis is about freedom from outside control. Though a robot may be sufficiently competent to perform a range of actions, it may be constrained from doing so by a variety of social and environmental factors. "Low" indicates that, although possibly capable of performing the task, the robot is not permitted to do so. "High" indicates the robot has the authority over its own actions, though it does not necessarily imply sufficient competence.

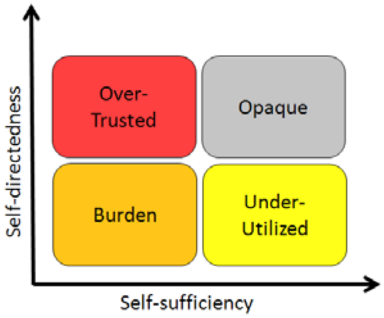


Fig. 3 Common system issues mapped against an autonomy-centered plane

Direct teleoperation, in which both self-sufficiency and self-directedness are absent, corresponds to the region labeled *Burden*. Increasing the self-directedness without a corresponding level of self-sufficiency will result in a system that is *over-trusted*, as shown in the upper left of the figure. Many systems fall in this category, including, for example, every entry in the DARPA robotic vehicle Grand Challenge that failed to complete the task. When autonomous capabilities are seen as insufficient, particularly in situations where the consequences of robot error may be disastrous, it is common for self-directedness to be limited. When the system self-directedness is reduced significantly below the potential of its capabilities the result is an *underutilized* system, as shown in the lower right corner of the figure. An example of this would be the first generations of Mars rovers which, due to the high cost of failure, were not trusted with autonomous action, but rather were subject to the decisions of a sizable team of NASA engineers. Here is the key point, however:

> *Even when self-directedness and self-sufficiency are reliable, matched appropriately to each other, and sufficient for the performance of the robot's individual tasks, human-robot teams engaged in consequential joint activity frequently encounter the potentially debilitating problem of opacity, meaning the inability for team members to maintain sufficient awareness of the state and actions of others to maintain effective team performance.*

The problem of *opacity* in robotics was highlighted recently by Stubbs [19] but had been previously identified as a general challenge more than two decades ago by Norman [20]. Norman cites numerous examples of opacity, most of which come from aviation where silent (opaque) automation has led to major accidents. This opacity often leads to what Woods calls "automation surprises" [21] that may result in catastrophe. An example is an autopilot that silently compensates for ice build-up on the airplane wings, while pilots remain unaware. Then, when the limits of control authority are reached and it can no longer compensate for extreme conditions, the automation simply turns off, forcing the pilots to try to recover from a very dangerous situation.

In the next section, we discuss the importance of interdependence in joint activity, and highlight opportunities for addressing it.

4 Interdependence

Coactive Design takes *interdependence* as the central organizing principle among people and agents working together in joint activity. Our sense of joint activity parallels that of Clark [22], who has described what happens in situations when one party does depends on what another party does (and vice-versa) over a sustained sequence of actions [23]. In such joint activity, we say that team members are "interdependent."

In his seminal 1967 book, James D. Thompson [24] recognized the importance of interdependence in organizational design. He also noted that there was a lack of understanding about interdependence. Similarly, we feel that understanding interdependence is critical to the design of human-agent systems. Understanding the nature of the interdependencies involved provides insight into the kinds of coordination that will be required among groups of humans and agents. Indeed, we assert that coordination mechanisms in skilled teams arise largely because of such interdependencies [25]. For this reason, understanding interdependence is an important requirement in designing agents that will be required to work as part of human-agent systems engaged in joint activity. Below, we introduce three new concepts that are important extensions to previous work on interdependence, particularly in the context of Coactive Design of human-agent systems.

4.1 Hard vs. Soft Interdependence

In their interdisciplinary study of coordination, Malone and Crowston [26] summarized prior work on coordination from many fields. Like us, they view coordination as required for managing dependencies (though we would say interdependencies—more on that below). They also characterize some of the most common types of dependencies, e.g., use of shared resources, producer/consumer relationships, simultaneity of processes, and task/subtask roles. These types of dependencies have received considerable attention in the literature. Unfortunately, they are insufficient to capture the necessary types of interdependence in human-agent systems.

In his research, Malone specifically was concerned with dependency as a matter of understanding how the results of one task enable the performance of another. However, in joint activity, we are not exclusively interested in the hard constraints that enable or prevent the possibility of an activity, but also in the idea of "soft interdependence," which includes a wide range of "helpful" things that a participant may do to facilitate team performance. The difference between strict dependence and soft interdependence is illustrated in the contrast between the two situations shown in figure 4—one in which a train car is completely dependent on the engine to pull it, and the other in which two friends provide mutual support of a helpful nature that is optional and opportunistic rather than strictly required. Indeed, our observations to date suggest that good teams can often be distinguished from great ones by how well they support requirements arising from soft interdependencies.

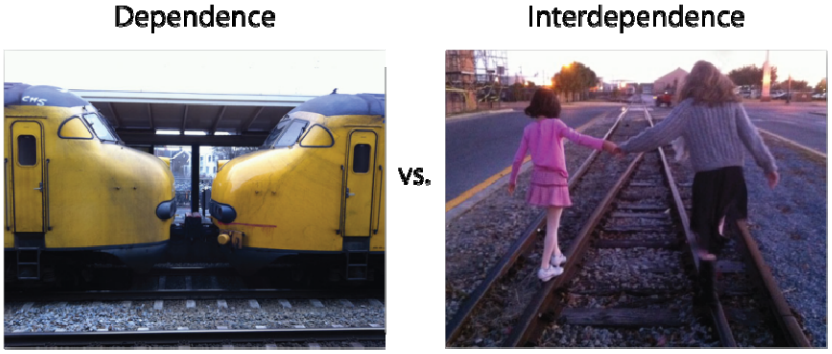


Fig. 4 Dependence vs. Interdependence

Examples of such forms of interdependence often seen among effective human teams include progress appraisals [27] ("I'm running late"), warnings ("Watch your step"), helpful adjuncts ("Do you want me to pick up your prescription when I go by the drug store?"), and observations about relevant unexpected events ("It has started to rain"). They can also be physical actions, such as opening a door for someone who has their hands full. Though social science research on teamwork clearly demonstrates their importance, soft interdependencies have been relatively neglected by agent researchers.

Although some previous human-agent systems have succeeded in supporting various aspects of teamwork that relate to soft interdependence, they have often lacked convincing general principles relating to their success. We are hopeful that the concept of interdependence can eventually provide such principles. In the meantime, we have at least become convinced that human-agent systems defined solely in terms of traditional notions of hard dependence and autonomy limit the potential for effective teamwork, as the preliminary experimental results discussed in a later section seem to indicate.

4.2 Inter-Activity Dependence vs. Intra-Activity Interdependence

Thompson [24] suggested three types of interdependence: *pooled, sequential and reciprocal*. Pooled interdependence describes a situation in which each entity contributing (independently) a discrete part to the whole, with each in turn being supported by the whole. Sequential interdependence occurs when one entity directly depends on the output of another—to us this would be better described as simple *dependence*. Reciprocal interdependence is a bidirectional sequential interdependence or what we would call mutual dependence.

Thompson's three types of interdependence are described in terms of how the output or product of an entity affects other entities engaged in independent activities. They do not, however, adequately model the full range of interdependencies involved in joint activity. Thompson's types can be viewed as *inter*-activity dependence. For human-agent systems engaged in joint activity there remain other types that can be considered *intra*-activity interdependence. For example, progress appraisal (determining and sharing with others how one's task "is going") and notifying others of unexpected events [27] are usually performed *within* an ongoing activity. We will call this *supportive interdependence*. In future research, this type of interdependence will be further elaborated, and additional types of interdependence in joint activity will be identified.

4.3 Monitoring as a Requirement for Handling Supportive Interdependence

The problem of monitoring for conditions that relate directly to an assigned agent task, apart from the vagaries of sensing itself, presents a few challenges for agent developers. If, for example, an agent needs an elevator (resource dependence), the agent can monitor the elevator doors to see when they open. Alternatively, the agent could be notified of availability (sequential interdependence) through signaling (e.g. up arrow light turns on, audible bell, or an elevator operator telling you "going up").

However, handling supportive interdependence often requires groups of agents and people to monitor the ongoing situation, to "look out for each other," even when the aspects of the situation being monitored do not relate directly to a given individual's assigned tasks. For example, in order to provide back-up behavior to compensate for a teammate's frail self-sufficiency, other team members might decide to monitor the teammate to know when it is appropriate to provide assistance. Monitoring interdependence also highlights the reciprocal nature of the activity. Not only does the monitoring entity need to monitor, but the monitored entity may need to make certain aspects of its state and behavior observable.

5 Coactive Design

The fundamental principle of Coactive Design is that interdependence must shape autonomy. Certainly joint activity of any consequence requires a measure of autonomy (both self-sufficiency and self-directedness) of its participants. Without a minimum level of autonomy, an agent will simply be a burden on a team, as noted by Stubbs [19]. However, it can be shown that in some situations simply adding more autonomy can hinder rather than help team performance. The means by which that agent realizes the necessary capabilities of self-sufficiency and self-directedness must be guided by an understanding of the interdependence between team members in the types of joint activity in which it will be involved. This understanding of interdependence can be used to shape the design and implementation of the agent's autonomous capabilities, thus enabling appropriate interaction with people and other agents.

In contrast to autonomous systems designed to take humans out of the loop, we are specifically designing systems to address requirements that allow close and continuous interaction with people. As we try to design more sophisticated systems, we move along a maturity continuum [28] from dependence to independence to interdependence. The process is a continuum because at least some level of independence of agents through autonomous capabilities is a prerequisite for interdependence. However, independence is not the supreme achievement in human-human interaction [28], nor should it be in human-agent systems. Imagine a completely capable autonomous human possessing no skills for coactivity—how well would such a person fit in most everyday situations?

The dictionary gives three meanings [29] to the word "coactive": 1) Joint action, 2) An impelling or restraining force; a compulsion, 3) Ecology; any of the reciprocal actions or effects, such as symbiosis, that can occur in a community. These three meanings capture the essence of our approach and we translate these below to identify the three minimum requirements of a coactive system. Our contention is that for an agent to effectively engage in joint activity, it must at a minimum have:

1) *Awareness of interdependence in joint activity*
2) *Consideration for interdependence in joint activity*
3) *Capability to support interdependence in joint activity*

We are not suggesting that all team members must be fully aware of the entire scope of the activity, but they must be aware of the interdependence in the activity. Similarly, all team members do not need to be equally capable, but they do need to be capable of supporting their particular points of interdependence. We now address each requirement in more detail.

5.1 Awareness of Interdependence in Joint Activity

In human-machine systems like today's flight automation systems, there is a shared responsibility between the humans and machines, yet the automation is completely unaware of the human participants in the activity. Joint activity implies mutual engagement in a process extended in space and time [22, 30]. Previous work in human-agent interaction has focused largely on assigning or allocating tasks to agents that may know little about the overall goal of the activity or about other tasks on which its tasks may be interdependent. However, the increasing sophistication of human-machine systems depends on a mature understanding of the requirements of interdependence between team members in joint activity.

Consider the history of research and development in unmanned aerial vehicles (UAVs). The first goal in its development was a standard engineering challenge to make the UAV self-sufficient for some tasks (e.g., stable flight, waypoint following). As the capabilities and robustness increased, the focus shifted to the problem of self-directedness (e.g., what am I willing to let the UAV do autonomously). The future directions of UAVs indicate a another shift, as discussed in the Unmanned Systems Roadmap [31] which states that unmanned systems "will quickly evolve to the point where various classes of unmanned systems operate together in a cooperative and collaborative manner…" This suggests a need to focus on interdependence (e.g., how can I get multiple UAVs to work effectively as a team with their operators?). This pattern of development is a natural maturation process that applies to any form of sophisticated automation. While awareness of interdependence was not critical to the initial stages of UAV development, it becomes an essential factor in the realization of a system's full potential. We are no longer dealing with individual *autonomous* actions but with group *participatory* actions [22]. This is a departure from the previous approaches discussed in section 3, with the exception of Collaborative Control [5], which aimed to incorporate all parties into the activity through shared human-agent participation in perceptual and cognitive actions.

5.2 Consideration for Interdependence in Joint Activity

Awareness of interdependence is only helpful if requirements for interdependence are taken into account in the design of an agent's autonomous capabilities. As Clark states, "a person's processes may be very different in individual and joint actions even when they appear identical" [22]. One of Clark's favorite examples is playing the same piece of music as a musical solo versus a duet. Although the music is the same, the processes involved are very different. This is a drastic shift for many autonomous robots, most of which were designed to do things as independently as possible.

In addition to the processes involved being different, joint activity is inherently more constraining than independent activity. Joint activity may require participating parties to assume collective obligations [32] that come into play even when they are not currently "assigned" to an ongoing task. These obligations may require the performance of certain duties that facilitate good teamwork or they may limit our individual actions for the good of the whole. For example, we may be compelled to provide help in certain situations, while at the same time being prevented from hogging more than our share of limited resources. In joint activity, individual participants share an obligation to coordinate; sacrificing to a degree their individual autonomy in the service of progress toward group goals. These obligations should not be viewed as only a burden. While it is true they usually have a cost, they also provide an opportunity.

5.3 Capability to Support Interdependence in Joint Activity

While consideration is about the deliberative or cognitive processes, there is also an essential functional requirement. We have described self-sufficiency as the capability to take care of one's self. Here we are talking about the capability to support interdependence. This means the capability to assist another or be assisted by another. The coactive nature of joint activity means that there is a reciprocal requirement in order for interdependence to be supported, or to put it another way, there is the need for complementary capabilities of those engaged in a participatory action. For example, if I need to know your status, you must be able to provide status updates. If you can help me make navigation decisions, my navigation algorithm must allow for outside guidance. Simply stated, one can only give if the others can take and vice versa. The abilities required for good teamwork require reciprocal abilities from the participating team members.

6 Visualizing the New Perspective

So how does the coactive design perspective change the way we see the agent design problem? In section 3.6, we depicted the two senses of autonomy on two orthogonal axes representing an autonomy-centered plane of agent capabilities. Coactive Design adds a third orthogonal dimension of agent capability: support for interdependence (figure 5).

The *support for interdependence* axis characterizes an agent in terms of its capability to depend on others or be depended on by others in any of the dimensions of autonomy. This axis is specifically about the capability to be interdependent, *not* the need or requirement to *be* dependent which are captured by the other axes. Although we are showing a single set of axes for simplicity, The Coactive Design perspective considers all dimensions [8] as discussed in section 3.5. The take away message is not the support of any particular cognitive model, but instead the concept that there are many aspects to an agent as it performs in a joint activity. Just as Castelfranchi argued that autonomy can occur at any of these "levels" or dimensions, Coactive Design argues that the ability to be *interdependent* exists at each "level" or dimension as well.

As we look at the challenges faced by current autonomous systems from a Coactive Design perspective, we see not only the constraints imposed by interdependence in the system, but also as a tremendous opportunity. Instead of considering the activity an independent one we can think about it as a participatory [22] one. Both the human and the machine *are* typically engaged in the *same* activity. There may be domains where we would like a robot to go on its mission and simply return with a result, but most domains are not like this. We need the agent to have some self-sufficiency and self-directedness, but we remain interdependent as the participatory task unfolds. Supporting this need provides an opportunity to address some of the current challenges. Figure 5 lists just a few such opportunities. For example, over-trusted robots can be supplemented with human assistance and opaque systems can provide feedback and transparency. In fact, many of the ten challenges [2] of automation, such as predictability and directability apply to this new dimension.

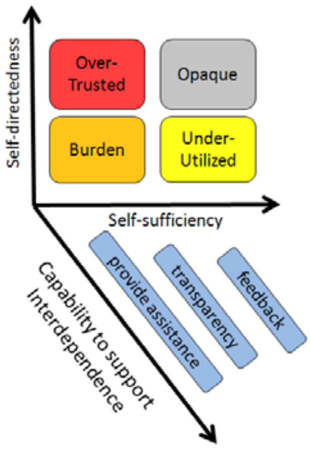


Figure 5 Support for interdependence as an orthogonal dimension to autonomy and some opportunities this dimension offers

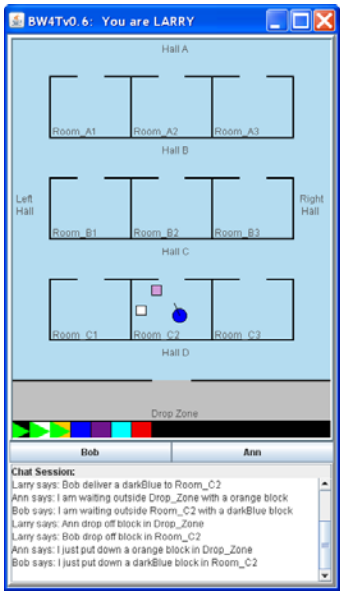
We can now map examples of prior work in autonomy onto this space (table 1). Section 3 describes how previous work was focused on self-sufficiency and self-directedness. Coactive Design presents the unique perspective of the *support for interdependence* dimension which is captured in the two rightmost columns of Table 1: the ability to depend on others and the ability to be depended on by others. The most important innovation of the Collaborative Control [5] approach was in accommodating a role for the human in providing assistance to the robot at the perceptual and cognitive levels. In other words, the robot had the ability to depend on the human for assistance in perception. The key insight of Collaborative Control was that tasks may sometimes be done more effectively if performed jointly. Coactive Design extends this perspective by providing a complement of this type of interdependence, accommodating the possibility of machines assisting people.

Table 1 Scope of concerns addressed by different approaches.

Approach	Autonomy-Centered		Teamwork-Centered (Support for Interdependence)	
	Self-sufficiency	Self-directedness	Ability to depend on others	Ability to be depended on
Functional Allocation	🟩			
Supervisory Control	🟩			
Adjustable Autonomy	🟩	🟩		
Sliding Autonomy	🟩	🟩		
Adaptive Autonomy	🟩	🟩		
Flexible Autonomy	🟩	🟩		
Mixed Initiative Interaction	🟩	🟩		
Collaborative Control	🟩	🟩	🟩	
Coactive Design	🟩	🟩	🟩	🟩

7 Initial Experiments

We have begun a series of experiments that relate to the fundamental principle of Coactive Design. Our first domain, Blocks World for Teams (BW4T) [33] was designed to be as simple as possible.



Chat Session:
Larry says: Bob deliver a darkBlue to Room_C2
Ann says: I am waiting outside Drop_Zone with a orange block
Bob says: I am waiting outside Room_C2 with a darkBlue block
Larry says: Ann drop off block in Drop_Zone
Larry says: Bob drop off block in Room_C2
Ann says: I just put down a orange block in Drop_Zone
Bob says: I just put down a darkBlue block in Room_C2

.

Fig. 6 BW4T game interface

Similar in spirit to the classic AI planning problem of Blocks World, the goal of BW4T is to "stack" colored blocks in a particular order. To keep things simple, the blocks are unstacked to begin with, so unstacking is not necessary. The most important variation on the problem we have made is to allow multiple players to work jointly on the same task. We control the observability between players and the environment. The degree of interdependence that is embedded in the task is represented by the complexity of color orderings within the goal stack. The task environment (figure 6) is composed of nine rooms containing a random assortment of blocks and a drop off area for the goal. The environment is hidden from each of the players, except for the contents of the current room. Teams may be composed of two or more players, each working toward the shared team goal. Players cannot see each other, so coordination must be explicit through the chat window. The task can be done without any coordination, but it is clear that coordination (i.e., the players managing their interdependence) can be beneficial.

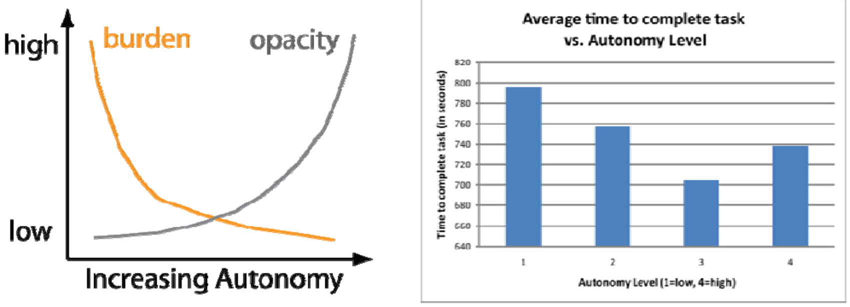
7.1 Added Autonomy without Addressing Interdependence Reduces Performance

A common suggestion for how to improve the performance of human-agent systems is to increase the level of agent autonomy [34, 35]. This solution is also commonly proposed for future systems [31]. It is true that additional increments of agent autonomy *might*, in a given circumstance, reap benefits to team performance through reduction of human burden. *However, there is a point in problem complexity at which the benefits of autonomy may be outweighed by the increase in system opacity when interdependence issues are not adequately addressed.* The fundamental principle of Coactive Design is that, in sophisticated human-agent systems, the underlying interdependence of participants in joint activity is a critical factor in human-agent system design. Another way to state this is that in human-agent systems engaged in joint activity, the benefits of higher levels of autonomy cannot be realized without addressing interdependence through coordination. Initial experiments using our BW4T domain seem to provide evidence for this claim.

For this experiment, we had a single human participate in a joint activity (collecting colored blocks in a specified sequence) with a single agent player. Both the human and the agent controlled a robot avatar. The agent teammate was directed by the human at levels of autonomy that varied in each experimental condition. The agent was designed to perform reliably and with reasonably intelligent behavior. This means that the self-directedness is always sufficient for the self-sufficiency and thus the system cannot be over-trusted. This experiment also limited the command interface for each level to the highest possible command set, thus preventing under-utilization. As such, we were looking only at the burdensomeness and opacity of the system.

In our lowest level of autonomy, Level 1, the human made all decisions and initiated all actions for the agent player. In essence, the human was manually controlling two robot avatars. This corresponds to Sheridan's [17] lowest level of autonomy. For Level 2 we automated most actions of the agent. All decisions remained with the human. We expected this automation would improve performance because it was reducing burden without adding opacity. Level 3 had all of the autonomous actions from Level 2 and also added an autonomous decision (i.e., which room to search). This increased opacity in two ways. First, the human is no longer aware of all of the decisions because one of them has been automated. Second, the robot has to make the decision without the same information the human had available when making the decision for the agent. Level 4 added automation of the remaining decision, making the task “fully autonomous.” This corresponds to Sheridan’s [17] highest level of autonomy.

We expected the burden to reduce from Level 1 to Level 4 and this was confirmed by an exit survey of the participants. However, we expected that as more activity and decision-making were delegated to the agent there would be an increased opacity in the system, reflected in more difficulties in the participants’ understanding of what was happening at a given moment. This was also confirmed by an exit survey of the participants. Finally, we expected this increase in opacity would result in decreased team performance. Time to complete the task was the performance metric, with lower times being better performance. The curves in figure 7a illustrate the general shape of results we expected, with the benefits of reduced human burden being eventually outweighed by the cost of opacity as autonomy increased beyond the inflection point.



a) Predicted results

b) Experimental results

Fig. 7 a) Hypothetical graph suggesting that the benefits of reduced human burden would eventually by outweighed by the cost of opacity as autonomy increases; b) Experimental results of 24 participants displayed as Average time to complete task vs. Autonomy level.

We ran an initial set of 24 subjects through all four levels (repeated measures) using a Latin Square design. While space prohibits a complete description of these first results here, our results were consistent with our prediction. Figure 7b shows the average times of all participants for each level. The predicted inflection point is apparent. While this is a single example in a single task domain, the results are consistent with the hypothesis that the benefits of higher levels of autonomy cannot be realized without addressing interdependence. If the general result holds as we continue our series of experiments, it will be a compelling demonstration of issues that cannot be addressed by autonomy-centered approaches, but can benefit from using the Coactive Design perspective.

7.2 Soft Interdependence Is a Key Factor in Performance

We have also run a pilot study of human-only teams to evaluate interdependence in the Block World for Teams domain. Although a simple domain, it demonstrates the complexity of coordination and interdependence even in the simplest domain. We ran twelve subjects in various team sizes (2, 3, 4, 5, 6, and 8). The subjects were allowed to talk openly to one another. As the activity became more interdependent (more complex ordering of the goal stack), we noted an increase in the number of coordination attempts, as would be expected. We also noted some interesting aspects of the communication. Although only two basic tasks are involved, we observed a wide variety of communications. Of particular interest were the large number of communications that were about *soft interdependencies* and monitoring issues that were related to them. An example of a soft interdependency is the exchange of world state information. Since players could only see the status of their current room, they would exchange information about the location of specific colors. Although the task could clearly be completed without this communication, the importance of this soft interdependence is demonstrated by the frequency of its use. An example of monitoring in support of interdependence issues was when players provided or requested an update as a colored block was picked up. The frequency of both progress updates and world state updates are examples of the importance of addressing *supportive interdependence* in human-agent systems for joint activity. These types of exchanges typically accounted for approximately 60% of the overall communication and increased with the degree of interdependence required for a given problem. A final observation was that not only the amount of communication changed with the degree of interdependence in the task, but the pattern of communication varied as well. For example, during tasks with low interdependence, world state and task assignment were the dominant communications. As interdependence in the task (complexity in the ordering of the goal stack) increased, they both diminished in importance and progress updates became dominant.

8 Discussion

The target for research in Coactive Design is not to support the development of current teleoperated systems or systems struggling with basic self-sufficiency. We are specifically addressing what a human-agent system would look like if it were to fill the more challenging roles of the future. The envisioned roles, if properly performed, have a greater level of interdependence that cannot be addressed solely by adjusting who is in control or who is assigned what task—and necessitate a focus on the coactivity. In contrast to autonomous systems designed to take humans out of the loop, we are specifically addressing the requirements for close and continuous interaction with people. The fundamental principle of Coactive Design provides a new perspective for designers of human-agent systems and gives some initial high-level guidance about what considerations are important. We plan to extend and expand this initial fundamental principle in future work.

In our first experiment, we have tried to demonstrate the issues with taking an autonomy-centered approach. By identifying the interdependence in the system, we can understand that there is a potential inflection point for team performance as autonomy increases. Awareness of this effect and its cause can help designers address the interdependence and improve performance, thereby yielding the full potential from autonomous capabilities. We plan to demonstrate this in future experiments.

We deliberately used a single human and single agent in our first experiment to show that even in the simplest case, our claim is still valid. We expect the effects to be more dramatic in larger teams and teams with higher levels of interdependence. Our demonstration used simple task interdependence, but there are other sources of interdependence including the environment, the team structure, and the team member capabilities. Future work will include developing a better understanding of the different types of interdependence.

We also used perfect autonomy for our experiment to show that even under ideal conditions, our claim is still valid. In real world systems, perfect autonomy will continue to be an elusive goal. This underlying truth necessitates human involvement at some level and accentuates the importance of teamwork. Agent frailties means one will have unexpected events (failures). One cannot overcome failed autonomy with autonomy, but one can possibly do so with teamwork (e.g., Fong's collaborative control [5]). Additionally, Christofferson and Woods [36] describe the "substitution myth": the erroneous notion that automation activities simply can be substituted for human activities without otherwise affecting the operation of the system. Even if frailty were not an issue, the "substitution myth" reminds us that autonomy is not removing something, but merely changing the nature of it. Humans cannot simply offload tasks to the robots without incurring some coordination penalty. This is not a problem as long as we keep in mind that autonomy is not an end in itself, but rather a means to supporting productive interaction [18]. Coactive Design reminds us that interdependence can provide opportunities to counteract these costs.

As agents move toward greater and greater autonomy, several researchers have expressed concerns. Norman states that "the danger [of intelligent agents] comes when agents start wresting away control, doing things behind your back, making decisions on your behalf, taking actions and, in general, taking over [37]." Simply deciding who is doing what is insufficient, because the human will always need to understand a certain amount of the activity. Additionally, humans are typically the desired beneficiaries of the fruits of the robot labor. We are the reason for the system and will always want access to the system. Not only do we want access to understand the system, but we also want to have input to affect it. To paraphrase Kidd [38], it is not merely that human skill is required, but also that human involvement is desired. That involvement means the human-agent system is interdependent.

8 Summary

We have introduced *Coactive Design* as a new approach to address the increasingly sophisticated roles for people and agents in mixed human-agent systems. The fundamental principle of Coactive Design recognizes that the underlying *interdependence* of participants in joint activity is a critical factor in the design of human-agent systems. In order to enable appropriate interaction, an understanding of the potential interdependencies among groups of humans and agents working together in a given situation should be used to shape the way agent architectures and individual agent capabilities for autonomy are designed. We no longer look at the primary problem of the research community as simply trying to make agents more independent through their autonomy. Rather, in addition, we strive to make them more capable of sophisticated interdependent joint activity with people.

References

1. Bradshaw, J. M., Paul Feltovich, and Matthew Johnson. “Human-Agent Interaction.” In *Handbook of Human-Machine Interaction,* edited by Guy Boy, in press. Ashgate, 2011.
2. Klein, Gary, David D. Woods, J. M. Bradshaw, Robert Hoffman, and Paul Feltovich. “Ten challenges for making automation a “team player” in joint human-agent activity.” *IEEE Intelligent Systems 19*, no. 6 (November-December 2004): 91-95.
3. Allen, J.E., C.I. Guinn, and E. Horvtz, Mixed-Initiative Interaction. IEEE Intelligent Systems, 1999. 14(5): p. 14-23.
4. Kortenkamp, D., Designing an Architecture for Adjustably Autonomous Robot Teams, in Revised Papers from the PRICAI 2000 Workshop Reader, Four Workshops held at PRICAI 2000 on Advances in Artificial Intelligence. 2001, Springer-Verlag.
5. Fong, T.W., Collaborative Control: A Robot-Centric Model for Vehicle Teleoperation. 2001, Robotics Institute, Carnegie Mellon University: Pittsburgh, PA.
6. Brookshire, J., S. Singh, and R. Simmons. Preliminary Results in Sliding Autonomy for Coordinated Teams. in Proceedings of The 2004 Spring Symposium Series. 2004.
7. Bradshaw, Jeffrey M., Alessandro Acquisti, James Allen, Maggie R. Breedy, Larry Bunch, Nate Chambers, Paul Feltovich, Lucian Galescu, M. A. Goodrich, Renia Jeffers, Matthew Johnson, Hyuckchul Jung, James Lott, D. R. Olsen Jr., Maarten Sierhuis, Niranjan Suri, William Taysom, Gianluca Tonti, and Andrzej Uszok. "Teamwork-centered autonomy for extended human-agent interaction in space applications." Presented at the AAAI 2004 Spring Symposium, Stanford University, CA, 22-24 March, 2004.
8. Bradshaw, Jeffrey M., Paul Feltovich, Hyuckchul Jung, Shri Kulkarni, William Taysom, and Andrzej Uszok. "Dimensions of adjustable autonomy and mixed-initiative interaction." In *Agents and Computational Autonomy: Potential, Risks, and Solutions. Lecture Notes in Computer Science, Vol. 2969,* edited by Matthias Nickles, Michael Rovatsos and Gerhard Weiss, 17-39. Berlin, Germany: Springer-Verlag, 2004.
10. Fitts, P.M., Human engineering for an effective air-navigation and traffic-control system. 1951, Washington,: National Research Council, Division of Anthropology and Psychology, Committee on Aviation Psychology. xii, 84 p.
11. Sheridan, T.B., Telerobotics, automation, and human supervisory control. 1992, Cambridge, Mass.: MIT Press. xx, 393 p.
12. Dorais, G. and D. Kortenkamp, Designing Human-Centered Autonomous Agents, in Revised Papers from the PRICAI 2000 Workshop Reader, Four Workshops held at PRICAI 2000 on Advances in Artificial Intelligence. 2001, Springer-Verlag.
13. Dias, M.B., Kannan, B., Browning, B., Jones, E., Argall, B., Dias, M.F., Zinck, M.B., Veloso, M.M., and Stentz, A.T., Sliding Autonomy for Peer-To-Peer Human-Robot Teams. 2008, Robotics Institute: Pittsburgh, PA. Myers, K.L. and D.N. Morley. Directing Agent Communities: An Initial Framework. in Proceedings of the IJCAI Workshop on Autonomy, Delegation, and Control: Interacting with Autonomous Agents. 2001. Seattle, WA.
14. K. L. Myers and D. N. Morley, Human directability of agents, Proceedings of the 1st international conference on Knowledge capture, ACM, Victoria, British Columbia, Canada, 2001.
15. Murphy, R., Casper, J., Micire, M., and Hyams, J. (2000) Mixed-initiative Control of Multiple Heterogeneous Robots for USAR.
16. Yanco, H.A. and J.L. Drury. A Taxonomy for Human-Robot Interaction. in AAAI Fall Symposium on Human-Robot Interaction. 2002.

17. Parasuraman, R., T. Sheridan, and C. Wickens, A model for types and levels of human interaction with automation. Systems, Man and Cybernetics, Part A, IEEE Transactions on, 2000. 30(3): p. 286-297.
18. Goodrich, M.A. and A.C. Schultz, Human-robot interaction: a survey. Found. Trends Hum.-Comput. Interact., 2007. 1(3): p. 203-275.
19. Stubbs, K., P. Hinds, and D. Wettergreen, Autonomy and common ground in human-robot interaction: A field study. IEEE Intelligent Systems, 2007(Special Issue on Interacting with Autonomy): p. 42-50.
20. Norman, D.A., The "problem" of automation: Inappropriate feedback and interaction, not "over-automation", in Human factors in hazardous situations, D.E. Broadbent, A. Baddeley, and J.T. Reason, Editors. 1990, Oxford University Press. p. 585-593.
21. Woods, D.D. and N.B. Sarter, Automation Surprises, in Handbook of Human Factors & Ergonomics, G. Salvendy, Editor. 1997, Wiley.
22. Clark, H.H., Using language. 1996, Cambridge [England] ; New York: Cambridge University Press. xi, 432 p.
23. G. Klein, P. J. Feltovich, J. M. Bradshaw and D. D. Woods, Common Ground and Coordination in Joint Activity, in K. R. B. William B. Rouse, ed., Organizational Simulation, 2005, pp. 139-184.
24. Thompson, J.D., Organizations in action; social science bases of administrative theory. 1967, New York,: McGraw-Hill. xi, 192 p.
25. Feltovich, P.J., Bradshaw, J.M., Clancey, W.J., and Johnson, M., Toward an Ontology of Regulation: Socially-Based Support for Coordination in Human and Machine Joint Activity, in Engineering Societies in the Agents World VII G. O'Hare, et al., Editors. 2007, Springer: Heidelberg, Germany. p. 175-192.
26. Malone, T.W. and K. Crowston, The interdisciplinary study of coordination. ACM Comput. Surv., 1994. 26(1): p. 87-119.
27. Feltovich, P.J., Bradshaw, J.M., Clancey, W.J., Johnson, M., and Bunch, L., Progress Appraisal as a Challenging Element of Coordination in Human and Machine Joint Activity, in Engineering Societies in the Agents World VIII A. Artikis, et al., Editors. 2008, Springer: Heidelberg, Germany. p. 124-141.
28. Covey, S.R., The 7 Habits of Highly Effective People. 1989, New York: Free Press.
29. coaction, in http://dictionary.reference.com/browse/coactive.
30. Sierhuis, M., "It's not just goals all the way down" - "It's activities all the way down" in Engineering Societies in the Agents World VII, 7th International, Workshop, ESAW 2006. 2007: Dublin, Ireland.
31. Office of the Secretary of Defense, Unmanned Systems Roadmap, 2007-2032.
32. van Diggelen, J., Bradshaw, J. M., Johnson, M., Uszok, A., and Feltovich, P. (2009). Implementing collective oblications in human-agent teams using KAoS policies. Proceedings of Workshop on Coordination, Organization, Institutions and Norms (COIN), IEEE/ACM Conference on Autonomous Agents and Multi-Agent Systems, Budapest, Hungary, 12 May 2009.
33. Johnson, M., et al., Joint Activity Testbed: Blocks World for Teams (BW4T) in Engineering Societies in the Agents World X. 2009.
34. Bleicher, A., The Gulf Spill's Lessons for Robotics, in ieee spectrum special report. 2010.
35. Jean, G.V., Duty Aboard the Littoral Combat Ship: 'Grueling but Manageable' in National Defense. 2010.
36. Christoffersen, K. and D.D. Woods, How to Make Automated Systems Team Players. 2002.
37. Norman, D.A., The invisible computer : why good products can fail, the personal computer is so complex, and information appliances are the solution. 1998, Cambridge, Mass.: MIT Press. xii, 302 p.
38. Kidd, P.T., Design of human-centered robotic systems, in Human-Robot Interaction, M. Rahimi and W. Karwowski, Editors. 1992, Taylor & Francis. p. 225-241.