



An Evaluation of Situational Autonomy for Human-AI Collaboration in a Shared Workspace Setting

Vildan Salikutluk
vildan.salikutluk@tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

Janik Schöpfer
janik.schoepper@stud.tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

Franziska Herbert
franziska.herbert@stud.tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

Katrin Scheuermann
katrin.scheuermann@stud.tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

Eric Frodl
eric.frodl@gmail.com
Technical University of Darmstadt
Darmstadt, Germany

Dirk Balfanz
dirk.balfanz@tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

Frank Jäkel
frank.jaekel@tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

Dorothea Koert
dorothea.koert@tu-darmstadt.de
Technical University of Darmstadt
Darmstadt, Germany

ABSTRACT

Designing interactions for human-AI teams (HATs) can be challenging due to an AI agent's potential autonomy. Previous work suggests that higher autonomy does not always improve team performance, and situation-dependent autonomy adaptation might be beneficial. However, there is a lack of systematic empirical evaluations of such autonomy adaptation in human-AI interaction. Therefore, we propose a cooperative task in a simulated shared workspace to investigate effects of fixed levels of AI autonomy and situation-dependent autonomy adaptation on team performance and user satisfaction. We derive adaptation rules for AI autonomy from previous work and a pilot study. We implement these rule for our main experiment and find that team performance was best when humans collaborated with an agent adjusting its autonomy based on the situation. Additionally, users rated this agent highest in terms of perceived intelligence. From these results, we discuss the influence of varying autonomy degrees on HATs in shared workspaces.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI; User studies; Collaborative interaction**; • **Computing methodologies** → **Artificial intelligence; Cognitive science**.

KEYWORDS

Human-AI Interaction, Shared Workspace, AI Autonomy



This work is licensed under a Creative Commons Attribution-NoDerivs International 4.0 License.

CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0330-0/24/05
<https://doi.org/10.1145/3613904.3642564>

ACM Reference Format:

Vildan Salikutluk, Janik Schöpfer, Franziska Herbert, Katrin Scheuermann, Eric Frodl, Dirk Balfanz, Frank Jäkel, and Dorothea Koert. 2024. An Evaluation of Situational Autonomy for Human-AI Collaboration in a Shared Workspace Setting. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 17 pages. <https://doi.org/10.1145/3613904.3642564>

1 INTRODUCTION

Recent artificial intelligence (AI) systems and robots as their embodied form show great potential to support humans at work [4, 49, 52] and in everyday tasks [32, 36, 48, 49, 72, 81]. Nevertheless, improving task outcomes through human-AI collaboration is not trivial because it often depends on the specific application area as well as on the abilities and characteristics of each party [13]. Even though there are helpful guidelines for classical human-computer interaction (HCI) already, they need to be updated for designing interactions with AI systems [3, 35, 79, 80] such that human-AI teams (HATs) can solve problems synergistically and improve their performance together. One factor that distinguishes human-AI interaction from classical HCI is the higher degree of autonomy that AI systems can possibly have during cooperative tasks [63, 80].

While there are many possible definitions for (AI) autonomy, in this paper, we refer to autonomy of a technical system as its ability to make decisions and execute actions independently without the need for constant human input [54]. The AI is not able to change the overall task goal, and its autonomy is limited by permissions and obligations given by humans and by environmental and task constraints [10]. Based on its autonomy level, the AI system can execute actions and initiate interactions with its human partner in order to complete its (sub-)tasks within the team successfully. When humans design technical systems as tools, they commonly automate a very specific sub-task to achieve their overall goals more efficiently by using the system [14, 42, 60]. In such cases,

these systems have a particular and limited purpose but no autonomy within the task and the team in general. While this describes the low end of an autonomy spectrum, on its other end is a fully autonomous partner with whom the human collaborates towards a common goal. This can be compared to human-human interaction, e.g., when colleagues at work collaborate with each other as a team. Such collaboration within HATs depends on several factors, e.g., the team's structure and composition [56, 74], the communication within the team [56, 82], and the skill level of each partner [34]. Often, the interaction with AI agents falls somewhere in between the two ends of the spectrum [77], as illustrated in Figure 1.

While AI needs a relatively high level of autonomy to be helpful in complex settings [50], high(er) autonomy in systems does not necessarily increase team performance or is preferred by their human counterparts in every situation [26]. Previous work also indicated that the ability to slide along the autonomy scale and dynamically adapt autonomy levels is beneficial [23, 47, 63, 70] and desired [33]. Adjustable autonomy is often specified as a set of autonomy levels [47, 53, 63, 83] which an operator can switch manually [21, 39, 47, 53, 83]. There are also specific use cases where automatically adjusting autonomy levels already showed promising results, e.g., multi-agent systems without human interaction [70], unmanned aerial vehicle path-planning [47], military helicopter cockpits [11], settings where an operator remotely controls a robot in hazardous environments [63], simulation-based evaluations for a cleaning and an inventory scenario with a mobile manipulator [23], or a non-cooperative object inspection task [59]. However, in previous work, we identified a lack of empirical evaluations of situational adaptation of AI autonomy in cooperative shared workspace settings with real human users.

Therefore, we propose a simulated shared workspace setting in which we study the effects of different autonomy levels and automatically switching between them during the interaction within a HAT. Specifically, in a study with 50 participants, we investigate how situational autonomy adaptation influences overall team performance in comparison to fixed autonomy levels (RQ1) and how it impacts the human's perception of the AI teammate and the interaction with it (RQ2). Our approach aligns well with the framework for conducting research about HATs by Cooke et al. [20]. They propose to identify essential aspects of HATs in the given domain, develop task environments and measurement strategies, and finally conduct human experiments to transfer the empirically validated insights in order to develop AI agents as teammates.

In summary, the main contributions of this paper are: First, we provide a well-defined task definition and autonomy level design for a cooperative task in a shared workspace that can be used to investigate effects of an AI agent's autonomy in such settings. Second, we propose general criteria for autonomy adaptation in shared workspace settings and derive specific heuristic rules for AI autonomy adaptation for our task from these criteria. Third, by using a concrete implementation, we add to the theoretical concepts in previous work with an empirical investigation on the effects of fixed as well as dynamic autonomy levels on HATs. In particular, we study the effects of AI autonomy levels on team performance and user satisfaction in cooperative shared workspace tasks with real users.

2 RELATED WORK

In this section, we first present related work about the interaction and collaboration within human-AI teams (Section 2.1) and provide a short background on theory of mind models, which we propose as an important factor to implement situational adjustment of autonomy in cooperative shared workspace settings. Furthermore, while there is a large body of previous work defining autonomy and its possible levels for AI or robotic agents [1, 10, 16, 26, 54, 56, 57], we focus on the discussion of related work that aims to enable AI agents to automatically adjust their level of autonomy (Section 2.2). For a more detailed overview on a definition of autonomy that aligns well with how it is framed in our work, we refer the interested reader to [10, 54].

2.1 Interactive Human-AI Teams

When humans work with AI systems, there is the potential to improve productivity and overall results in different work domains [52, 75]. Sometimes, human-AI teams (HATs) can even outperform human-human teams [51], which seems to be the case when there is an interdependence in the HAT [56]. Nevertheless, designing successful human-AI interaction can be difficult [79, 80] even if first guidelines exist [3, 35]. Specifically, how the interaction of a HAT needs to be designed, how their performance is measured, and how successful they are can depend on the concrete application and many other factors such as the team structure [56, 74], the communication between the teammates [56, 82], and their skill levels [34]. It is often necessary for human and AI skills to be complementary to each other [38, 40, 67]. In addition, teammates should have an awareness of the situation [22, 41] and about what their teammate knows and plans. This capability is known as theory of mind, i.e., the modeling of mental states of others [76]. These models are also used computationally in various human-AI interaction settings [7, 17, 23, 30, 37] and have been shown to influence team success for human-AI interaction [30]. Such models can ensure that systems adjust to specific users and plan better (together) with them [23, 45]. So while there are different ways on deciding how to delegate tasks within a HAT, e.g., by setting specific rules [43], theory of mind models (ToMMs) can help to improve the understanding about which task is suitable for which teammate [58] and distribute tasks better among them [34] to enable more efficient coordination.

2.2 Adaptive Autonomy

Dynamic adjustment of autonomy has been explored well in settings where multiple AI agents collaborate [31, 70]. Additionally, automatic adjustment of an AI agent's autonomy has been beneficial in settings where a human operator remotely controls one or multiple robots in hazardous or space environments [12, 25], has been used to adapt to different levels of user expertise [45], to control the amount of requested human input based on a robot's context-dependent self-confidence [59, 61], to adapt to the mental workload of the crew in a cockpit [11], and for path planning of unmanned aerial vehicles [47]. In particular, Lin and Goodrich [47] show that human-AI cooperation with autonomy adaptation can lead to better performance as opposed to either human or system completing the task alone. Generally, there is evidence that humans benefit from [69] and are in favor of systems having a high(er) level

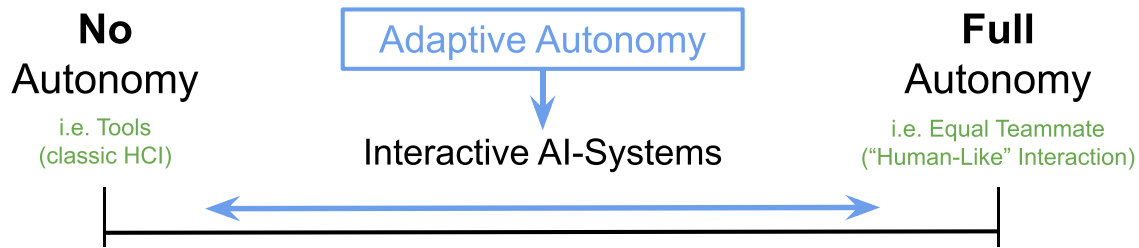


Figure 1: Overview of sliding degree of AI autonomy. It can range from no autonomy (left), i.e., like a tool completely controlled by its user, to fully autonomous (right) where a technical system becomes a somewhat equal teammate for the human. It is important to investigate the interaction paradigms arising when interactive AI systems slide along this autonomy scale. We propose to investigate the situational adaptation of autonomy in cooperative shared workspace settings.

of autonomy [63] when it helps them achieve their goals. Humans also share their task load more when they perceive a system’s behavior as human-like [73]. Nevertheless, there is also literature about how humans sometimes prefer when they have control over systems [3, 46] or reduce the systems’ autonomy [69]. Which level of autonomy users choose when they can switch manually between autonomy levels was investigated by Alan et al. [2] for AI-support in a task where users had to switch between electricity tariffs. In this study, only half of the participants adjusted the level to semi-autonomous. All other subjects kept the lowest level of autonomy, and none of the subjects chose the fully autonomous setting. An interview-based study on what autonomy level humans prefer in human AI-interaction for cyber incident response was presented by Hauptman et al. [33] and for instance showed that humans prefer higher autonomy in low risk settings and less autonomy and more control in high-stakes situations. Ball and Callaghan [6] trained an intelligent work space on human input to automatically adjust between four autonomy levels.

However, only few works evaluate the potential of adaptive autonomy in cooperative shared workspace settings [23, 29] and did not evaluate them in real user studies. Fiore et al. [29] present a framework that incorporates situation assessment and planning together with human intention prediction and reactive action execution. Their approach enables a robot to adapt to user preferences, allowing the human partner to be more passive or active in giving commands. A theory of mind model for predicting temporary absence or inattention of the human is proposed by Devin and Alami [23] to automatically adapt robot communication patterns during the execution of a cooperative table cleaning task. However, both approaches are only evaluated with simulated humans.

3 SITUATIONAL ADAPTIVE AUTONOMY FOR COOPERATIVE TASKS IN SHARED WORKSPACES

Efficient cooperation in human-AI teams within shared workspace settings is highly relevant in various application areas such as assisted living [18], industrial automation [28], or in assistive cockpit systems [11]. However, as discussed before, there is a lack of concrete implementations and systematic evaluations of situational autonomy adaptation for human-AI interaction in such settings.

While there are helpful (theoretical) conceptualizations for adaptive AI autonomy, it is crucial to empirically evaluate them with humans to ensure a validated human-centered approach for the development of future AI systems, which was not the focus of previous work. In Section 3.1 we formally define the setting that we consider here. Subsequently, we explain our choice of autonomy levels in this scenario in Section 3.2 and how we adapt them depending on situational features in Section 3.3.

3.1 Task Setup and Formalization of the Collaborative Shared Workspace Scenario

For our experimental evaluations, we implement a cooperative shared workspace setting as an office environment. Here, we focus on tasks that are purposely designed such that they can only be completed successfully with a collaboration between human and AI. The collaboration is set within a computer game. The AI is represented as a robot with a designated space separated by a counter from the space accessible to the human, as depicted in Figure 2. This task is inspired by the game *Overcooked*¹ and the environment is based on work by Rother et al. [62]. Similar task setups inspired by *Overcooked* have recently been used to investigate human-AI interactions (e.g., [9, 15, 44, 78]). In our office setting, the joint task for the human user and the agent is to process incoming boxes. For example, incoming documents need to be labeled and archived. Only the human can pick up the documents and only the agent can pick up the labels, hence, the two have to cooperate. There are other tasks as well that both can work on without the help of the other, but these need to be coordinated appropriately. We assume there is no debating or adjusting of the task goal. The goal is to be achieved with the help of the AI teammate [53] and the goal is known to both. We formalize this task similar to a Markov Decision Process (MDP) [8] where in each time step t in a state s_t the agent chooses an action a_t and the human chooses an action a_t^h resulting in an overall reward r_t . In the following subsections, we specify the set of possible actions, states, and the overall team reward.

3.1.1 Actions. The proposed task setting is intentionally designed such that there are actions that only the human or only the agent can execute, and actions that are feasible for both. However, one

¹<https://www.team17.com/games/overcooked/>

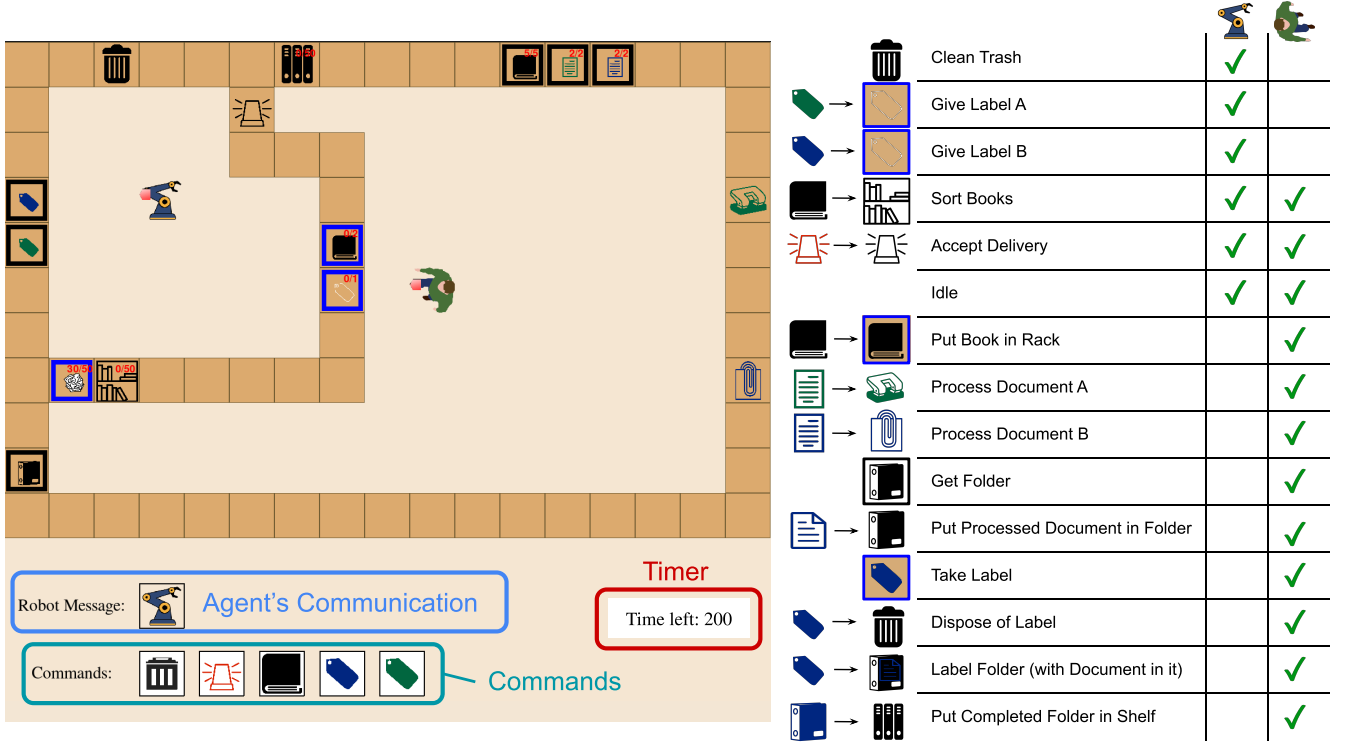


Figure 2: We implemented the cooperative shared workspace setting as an office environment in which a human and an AI agent (depicted as a robot) process and sort incoming document and book deliveries together. The human can control the human avatar and interact with objects over the keyboard, give commands to the agent (green box left), see agent’s messages (blue box) and a timer (red). The list of object actions for the agent and possible human actions is shown on the right.

of them might outperform the other, e.g., in terms of execution speed. For the agent, we distinguish between object actions A^o and communication actions A^c . This results in the set of overall possible actions for the agent

$$A = \{A^o, A^c\} = \{a_1^o, a_2^o, \dots, a_N^o, a_1^c, \dots, a_2^c, \dots, a_M^c\}, \quad (1)$$

where N denotes the number of possible object actions and M denotes the number of communication actions. The list on the right in Figure 2 summarizes the object actions for the agent and the executable actions for the human in the setting used for our experimental evaluation. More concretely, object actions formalize the interaction with the task environment, i.e., moving towards or processing objects, such as SORT-BOOK or GIVE-LABEL-A. Communication actions are used to inform or ask the human about a change to the agent’s current action or suggest a change to the human’s action. The set of possible communication actions of the agent towards the human is defined as $A^c = \{\text{Present all feasible options of } A^o \text{ in } s_t \text{ and ask human to choose; Present agent’s current perceived best action } a_t^* \text{ in } s_t \text{ and ask for confirmation; Switch to execute agent’s current perceived best action } a_t^* \text{ in } s_t \text{ and inform human about action switch; Inform human about an event and suggest to the human to execute a specific action } a_h \in A^h\}$.

3.1.2 States. We distinguish between the true underlying environment state and the environment state as the agent currently perceives it s_t . Only the latter can be used by the agent to make choices about actions or situational autonomy adaptation. The agent has full access to its own current state s_t^r , however parts of the human state s_t^h or object states s_t^o might be only partially observable and the agent can only form a belief about the true underlying states over time. As a result, the state from the agent’s viewpoint is defined as

$$s_t = \{s_t^r, b(s_t^h), b(s_t^o)\}, \quad (2)$$

where $b(s_t^h)$, denotes the agent’s current belief over the human state and $b(s_t^o)$ denotes the agent’s belief over the current object states. The state of the agent is defined as $s^r = \{x_r, y_r, \gamma_r, g_r\}$, where x_r and y_r denote the agent’s position in pixel coordinates, γ_r denotes its orientation and g_r the agent’s current action goal, e.g., SORT-BOOK or GIVE-LABEL-A. The state of the human is defined as $s^h = \{x_h, y_h, \gamma_h, g_h, \{o_1^{\text{fov}}, \dots, o_L^{\text{fov}}\}\}$, where x_h and y_h define the human position, γ_h denotes the human’s orientation, g_h denotes the current human action goal and $\{o_1^{\text{fov}}, \dots, o_L^{\text{fov}}\}$ are all objects that are currently within the human’s field-of-view. The object states are defined for each object j by their position x_o^j, y_o^j and their specific (boolean) properties p_j (e.g., processing state of the document, or ringing/not ringing for the doorbell).

For our experimental evaluation, we assume that the human position as well as all object positions and properties are fully observable for the agent, and it only needs to form a model about the current human goal and the human’s perception of objects. Specifically, the agent could compute an estimate about the next human goal based on the history of state-action pairs. For the field-of-view of the human in our task, the agent assumes the human can see straight ahead and 45 degrees of their periphery. This assumption matches the implemented area that is visible to the human during the task, as illustrated in Figure 3. All objects within this area are assumed to be visible to the human. However, it should be noted that due to human attention, the true perception of the human might differ from this assumption.

3.1.3 Overall Task Goal and Rewards. We consider a task setting where a human and an agent work together to maximize a joint team reward that is known to both of them. In our experimental implementation, the overall task goal for the human-AI team is to organize the contents of as many delivery boxes as possible within a fixed time limit. The boxes contain two different types of documents (green and blue) and books. The documents have to be processed, filed into correctly labeled folders, and stored in a shelf. There are also books that need to be placed in a bookshelf. Additionally, trash needs to be thrown into the bin. In each time step, we formally define the team reward r_t as

$$r_t := \begin{cases} +5, & \text{if completed folder is sorted into shelf;} \\ +5, & \text{if book is sorted into shelf;} \\ +1, & \text{if trash is put into trash bin.} \end{cases}$$

At the beginning of the task, there is one box in the shared workspace. New boxes are delivered one after another over time. To maximize team reward, it is crucial that not only are the documents and books in the boxes organized correctly, but also that as many box deliveries are accepted as possible. A new delivery is always indicated by a ring of the doorbell (by the doorbell symbol turning from black to red and a timer being displayed in it). Either the human or the agent have to answer the doorbell for the box to be delivered, but due to the human’s limited field-of-view (as shown in Figure 3 the participants cannot always see the doorbell (and therefore might miss that it is ringing). As shown in Figure 2, the boxes can contain up to five books and two of each document type. When a delivery is accepted, the boxes get filled to their limit. Thus, the aim is to work through as much of the content as possible until the next delivery arrives. For example, if all five books are sorted into the shelf, five new books can be delivered next. However, if only two are sorted, then only two can be delivered (as the maximum limit remains five). If the doorbell is not tended to within a defined time limit, the box delivery is missed.

3.2 AI Autonomy Levels and Action Selection

To investigate effects of situational adaptive autonomy on human-AI interaction in a shared workspace, we require a concrete implementation of autonomy levels for the AI agent. In Section 3.2.1 we explain our design of autonomy levels and compare it to existing definitions in the literature. Subsequently, in Section 3.2.2 we describe the influence of the AI agent’s autonomy level on its action selection process.

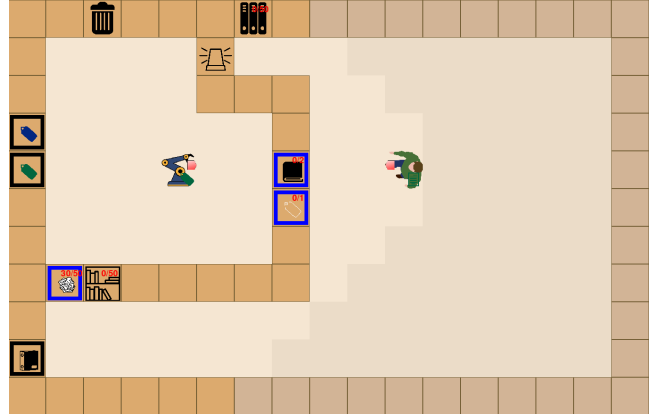


Figure 3: The task setup is displayed here with the field-of-view that is activated during the experiment. During the trials, participants cannot see, e.g., what is behind them as it is not shown with the limited field-of-view.

3.2.1 Autonomy Level Design. Existing conceptualizations for autonomy levels are often proposed abstractly and without concrete applications or tasks in mind [26, 27, 54, 56, 57, 64, 65]. Some conceptualizations focus on the level of automation that is technically possible [10] others on how much the human is involved in the decision-making process [26, 59], or on trying to define autonomy [57]. One of the most established conceptualizations is the model of Parasuraman et al. [56, 57]. Their model distinguishes 10 discrete levels of automation, where in level 1 the AI offers no assistance and in level 10 the AI acts fully autonomously, ignoring the human. From a practical point of view, if the AI was to adapt its autonomy level in different contexts, it might be challenging for humans to repeatedly adjust their mental model of the AI’s capabilities and possible interactions to 10 different autonomy levels [71]. Therefore, we decided to base our implementation of autonomy levels on a recent reduction of the original levels by Parasuraman et al. [57] into only three levels, i.e., *High Agent Autonomy*, *Partial Agent Autonomy* and *No Autonomy/Manual Control* which O’Neill et al. [56] introduced and Hauptman et al. [33] recently also used in their work on human-AI interaction for cyber incident response. We decided to split the *No Autonomy/Manual Control* level such that we distinguish between four autonomy levels. Specifically, in low autonomy the control is fully manual but in contrast to no autonomy the agent can actively ask for instructions when it assumes that there is a need to change its current action. Our resulting four autonomy levels can be categorized by how much initiative the AI agent takes within the human-AI team by either suggesting next actions once it completed a sub-task or alternatives to its assigned sub-task if it assumes a benefit for the overall team performance. Specifically, we distinguish between *No Autonomy*, *Low Autonomy*, *Moderate Autonomy* and *High Autonomy*.

Since we want to design a system that automatically adapts its autonomy level, we need to consider which situations would require (autonomous) action changes of the agent. Furthermore, these should be situations which typically arise during human-AI interactions. Table 1 summarizes the concrete implications of our

four autonomy levels on the agent’s behavior in three such situation types.

In *Situation Type 1* the agent encounters a problem that makes further execution of its current action infeasible. A concrete example of this situation type in our task is when during the SORT-BOOK action the agent notices that there are no more books left to sort within its reach. In this case, the agent can either switch to idle mode (no autonomy), ask the human what to do (low autonomy), suggest the next action it considers beneficial and wait for human confirmation (moderate autonomy) or directly execute the next action and inform the human about this switch (high autonomy).

In *Situation Type 2* the agent notices a higher priority task that makes a switch to another action within the agent’s own action space beneficial w.r.t. overall task performance. In our setting, this is the case when, e.g., the agent is sorting trash and notices that the doorbell is ringing. If the agent does not switch to accepting the delivery, it would be missed. The agent can either ignore the higher priority task and continue with its current action (no autonomy) or ask the human what to do (low autonomy). Alternatively, the agent could inform the human about the higher priority event, suggest the next action it considers best and wait for confirmation of the human (moderate autonomy), or the agent could directly switch to this action, i.e., autonomously accepting the delivery, and inform the human about its action switch (high autonomy).

Situation Type 3 is different from the first two in that it encompasses situations in which the agent notices that a change of not only its own action but also of the human’s action might lead to overall better task performance. For example, if the human sorts all books instead of placing them on the book rack, such that the agent can sort them, this can lead to reduced efficiency. Instead, the agent could ask the human to place the books on the rack. This way, the agent can contribute by sorting books and the human could instead process documents such that both teammates can work in parallel. In this type of situation, the agent can either make no suggestion for improved task distribution and just continue with its current action (no autonomy), ask the human what to do (low autonomy), or the agent suggests what it considers to be the best option for redistributing the tasks and waits for confirmation (moderate/high autonomy). Note that in this situation type, there is no difference between moderate and high autonomy, since a change in the human’s behavior always requires compliance of the human and therefore cannot be done fully autonomously by the agent.

3.2.2 Action Selection. The state s_t , which includes, e.g., the human’s position and estimated current action, the agent’s own relative position to objects, and current object properties, determines which actions are currently feasible for the agent. In case one of the three situation types as defined in Section 3.2.1 occurs, the agent determines which of its feasible actions it considers to be the next best action $a_{(t+1)}^*$. For our experimental evaluation, we choose a fixed, heuristic order for all possible sub-tasks. Based on this prioritization, the agent determines $a_{(t+1)}^*$. The order of priority from high to low is to answer the doorbell (if it cannot be answered by the human), provide labels if the agent recognizes a need for it from the human, sort books and lastly, dispose of trash.

If $a_{(t+1)}^*$ does not match the agent’s current action, it can execute one of its communication actions to initiate a switch of its action.

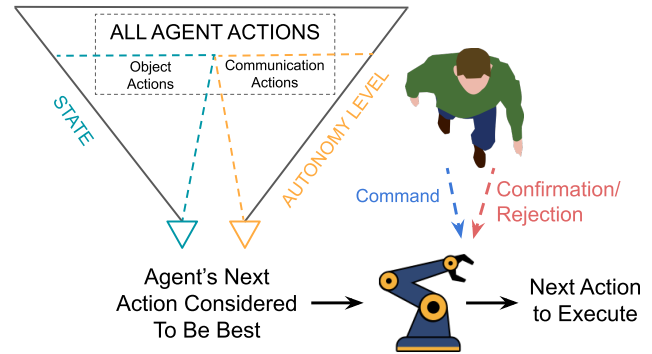


Figure 4: The agent selects its next action based on the current state and its autonomy level. Human commands overwrite the next planned action, even though the agent can question them once, depending on its degree of autonomy.

The agent’s current autonomy level determines which communication actions are available and how they are presented. In each time step, it is also possible for the human to command an action to the agent. However, in moderate, high or dynamic autonomy level the agent can suggest an alternative if it considers this action to not be the best one in the current state. Nevertheless, the agent will adhere to the given command of the human and only switch if the human confirms its alternative. This allows for the human to have control at all times, if necessary. Figure 4 summarizes how the agent decides on its next action.

3.3 Situational Autonomy Adaptation in Cooperative Shared Workspaces

While higher AI autonomy can be required for some complex tasks [50], it does not always increase the overall team performance [26]. Automatic adaptation of the AI agent’s autonomy level depending on the situation at hand might therefore be beneficial within human-AI teams [54, 70]. In this section, we propose that such situational autonomy adaptation for an AI agent in a cooperative shared workspace should be based on five criteria: 1) the agent’s self-confidence within a sub-task, 2) the effects of sub-task execution failure, 3) the agent’s theory of mind model about the human partner, 4) competence comparison between agent and human, and 5) whether a modification of human behavior is required or not.

The selection of these criteria and their implementation for our experimental setting is based on a literature review of existing related methods for autonomy adaptation and a pilot study. In this pilot study, we evaluated task performance with 28 participants who each completed 3 trials in one of the four fixed-autonomy levels that we defined in Section 3.2.1. We also collected additional think-aloud data with another 8 participants during 3 trials, again with fixed-autonomy levels (two participants per condition) which were transcribed and coded with a grounded theory approach [19, 68] with MAXQDA.² The setup differed slightly between the pilot study and our experiment presented in Section 4 as, e.g., there were fewer deliveries per trial, and we adapted the agent’s communication

²<https://www.maxqda.com/>

	Situation Type 1	Situation Type 2	Situation Type 3
	<i>Problem occurs during current action execution of the agent</i>	<i>During execution of its current action, agent notices higher priority event that could be addressed within its own action space</i>	<i>Agent notices benefit of a sub-task redistribution or higher priority event that influence the actions human should choose.</i>
No Autonomy	Idle	Ignore more important action. Continue with current action.	No suggestion for improved task distribution. Continue with current action.
Low Autonomy	Ask human what to do from the set of feasible actions in current state and wait for human commands.		
Moderate Autonomy	Suggest the alternative action considered best. Wait for human confirmation/rejection.	Inform the human about higher priority event and suggest alternative action considered best that the agent could switch to. Wait for human confirmation/rejection.	Suggest the option for task plan redistribution considered best between human and agent. Wait for confirmation/rejection.
High Autonomy	Execute the alternative action considered best. Inform human about action change.	Switch to best alternative action to address higher priority event. Inform human about action change.	

Table 1: Overview of situation types that can occur during our cooperative organizational task within our shared workspace setting and the agent’s behavior based on its autonomy level. The explanations show what the agent would recognize as the current type of situation and how it would act and communicate with its human teammate on each autonomy level based on the situation type.

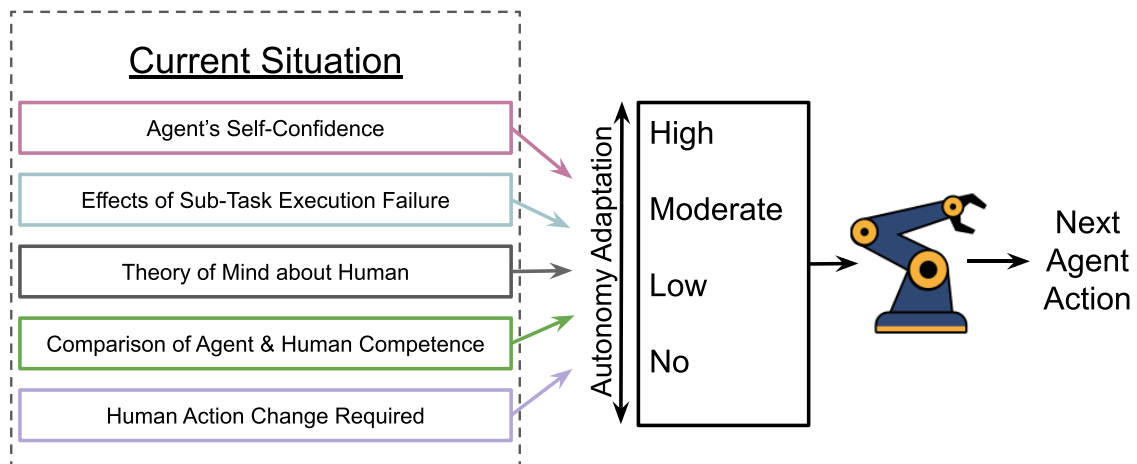


Figure 5: The adaptation of the AI agent’s autonomy is based on the current situation and our criteria for adjustments. These criteria are 1) the agent’s self-confidence within a sub-task, 2) the effects of sub-task execution failure, 3) the agent’s theory of mind model of the human partner, 4) competence comparison between agent and human, and 5) whether a modification of human behavior is required or not.

messages and their position based on the participants' feedback from the pilot study.

The first two criteria, i.e., the agent's self-confidence in its own competence to complete a certain sub-task correctly [29, 59, 61] and overall sub-task prioritization [29] are the most prominent of the five criteria in the literature. In particular, an AI system should lower its autonomy when uncertainties or problems arise or if a mismatch between its own competence and (sub-)task requirements occurs. An increase in autonomy can be advantageous, e.g., when it is evident that the agent can complete the task correctly and when only autonomously switching to a sub-task with higher priority can prevent catastrophic failures. In our pilot study, these two principles also held true for our experimental setting. Specifically, participants preferred a high degree of agent autonomy in tasks that were executed with high precision, i.e., sorting books when they were accessible to the agent or tending to the doorbell if it was needed. This was also reflected in the task performance as, e.g., in high autonomy conditions there were clearly more books sorted than in the other conditions.

However, in a cooperative shared workspace setting it is crucial to consider that the AI agent is not executing a task on its own but in a team with a human and thus needs to take its human partner into account [34, 56]. Therefore, theory of mind models (ToMMs), i.e., modelling the human's current and planned goals become an important aspect for the adaptation of the agent's autonomy level and with it its initiative and behavior. This has not yet been a main focus in the field of autonomy adaptation and was so far only considered in few works [23, 29, 45]. Nevertheless, Musick et al. [55] showed that the ability to predict the actions of teammates is central to performance in human-AI teams. This aligns well with insights from our pilot study, which showed that it was generally perceived well when the agent anticipated actions that were of benefit for the plan of the human teammate. If the agent, e.g., decided autonomously to hand over the needed label to the human, it was perceived positively.

However, there were also instances in the high-autonomy condition in our pilot study that demonstrated a need for the agent to consider its certainty in its model about the human partner. Specifically, when the agent predicted the wrong color for the required label and autonomously handed it over, it caused confusion and frustration for some participants. Two examples from our think-aloud data illustrate this well: One participant commented "There are two green documents lying here, and he wants to bring me a blue label"³, and another exclaimed "Again, weird that blue labels are suggested, even though I am just handling green documents all the time"³.

Another important aspect for dynamically switching the agent's autonomy is the comparison of competence between the teammates. For instance, in multi-agent settings where two or more AI agents work together, this is a basic underlying factor of task distribution and autonomy scaling within a team [29, 70]. However, it became clear in our pilot study that even if an agent correctly recognizes that the human could do a certain task better than itself and therefore suggests a redistribution of tasks to the human, the human is not always in favor of following the AI agent's suggestions. The

following quotes from participants illustrate this case, "[...] just now I find it annoying again that it is suggesting that. The agent should first finish its task."³ They also were irritated if the suggested change for their behavior was what the human already had in mind to do next: "Now again. The suggestion was not necessary because it was the only thing I could do anyway."³

We applied these five criteria which we identified based on the insights from previous work and our pilot study to implement a set of specific rules for the agent's autonomy adaptation in our task setting. These criteria, which are summarized in Figure 5, are important factors that we consider crucial to the interaction of successful human-AI teams. From these general criteria, we derived and implemented specific rules for the behavior of our AI agent in our task design. These rules, which are presented in Table 2, describe how the agent switches between different autonomy levels and, thus, behaves based on how it perceives the situation. For instance, our ToMM recognizes a set of pre-defined human actions from human position and object features and computes which objects are visible to the human from the estimated field of view to, e.g., predict that the human will need a label. Depending on the agent's self-confidence about which label color is currently needed, it can then determine which autonomy level to switch to. If the agent has low confidence, it switches to low autonomy and simply asks the human what to do (rule 1 in Table 2), but if its confidence is high, it switches to high autonomy and will directly bring the human the label of the predicted color (rule 3 in Table 2). It should be noted that these rules are only a first heuristic implementation of the proposed concept, and neither the criteria nor the derived rules are necessarily exhaustive. However, these rules offer a valuable starting point to gain empirical insights about the effect of situational autonomy adaptation in a cooperative shared workspace setting.

4 EXPERIMENTAL EVALUATION

In this section, we present our evaluation of team performance in human-AI collaboration in the simulated shared workspace that we presented in Section 3.1. In addition, we also report the participants' subjective perception of the collaboration with the AI agent.

4.1 Methods

We collected data from 50 participants (26 male, 24 female, 18–34 years old). The experiment was approved by the local ethics board and all participants provided informed consent. We use a between-subject design with five conditions (the four fixed autonomy levels, i.e., no, low, moderate, and high and the situational autonomy adaptation). Each participant completed three trials of eight minutes each. With 10 participants in each condition, we thus have 30 trials per condition overall. Each participant was instructed to solve the office task as best as possible together with their AI collaborator. The experiment was conducted in German, the participants' native language. All experiments were conducted in a lab setting at a desktop computer. In all conditions, the participants first received written instructions for the task, the goal, and the communication with the AI agent on screen. Additionally, participants completed two training trials to familiarize themselves with the environment, task, and controls. The participants controlled the human avatar with the arrow keys and interacted with objects using the space bar

³translated from German

Explanation of the Situation	Autonomy Level	Next Action
Agent recognizes the need for labels by the human; self-confidence of correct label recognition is low	Low	Ask human what to do
Agent recognizes the need for labels by the human; self-confidence of correct label recognition is medium	Moderate	Ask human if they want the proposed label and wait for confirmation/rejection
Agent recognizes the need for labels by the human; self-confidence of correct label recognition is high	High	Bring label of recognized color
While disposing of trash, the agent recognizes that books can be sorted	High	Sort Books
A delivery arrives and needs to be accepted, but the human does not recognize it or is too far away	High	Accept delivery
A delivery arrives and needs to be accepted, the human doing another task but is closer to doorbell than the agent	Moderate	Ask the human to answer the doorbell wait for confirmation/rejection if human rejects agent answers doorbell
Agent receives a (perceived suboptimal) command and makes a counter-proposal	Moderate	Ask human to switch to alternative suggested action and wait for confirmation/rejection; if rejected execute original human command
Agent idles e.g. because it completed its prior task and is confident about what next action to take	High	Execute Next Task
While sorting books, the agent runs out of books and the human is about to deliver new books	No	Go to Idle, (expecting to continue book sorting soon and do not disturb human with message)
Agent wants to change the perceived plan of the human (as human action is perceived as suboptimal)	Moderate	Ask human if they instead would execute the proposed action (that it is perceived to be more efficient); wait for confirmation/rejection

Table 2: Overview of concrete implemented rules for AI autonomy adaptation and the agent’s behavior for our cooperative organizational task within our shared workspace setting. The explanation of the situations describes what the agent would recognize as the current situation and adjust its autonomy level to the indicated ones in the middle column in order to derive its next action, which are shown on the right of the table.

on the keyboard. Commands to the agent were given by clicking on them with the mouse. We developed the environment, which is based on work by Rother et al. [62], with pygame (version 2.0.1).

The participants were informed that sorting books and processing folders was most relevant for the overall reward, while sorting trash was less relevant. We collected all game-based user inputs (commands, reactions to agent requests) and all of our objective measures (number of completed folders, number of sorted books, number of sorted trash) within the environment during the experimental trials using JSON files. Participants proceeded with the first training round of three minutes without the field-of-view activated, thus participants saw the entire setup as shown on the left in Figure 2. This allowed participants to familiarize themselves with the general workspace structure and game dynamics. All initial positions of the objects, the overall layout of the workspace and difficulty of the task itself remained identical throughout all training and experimental trials. Afterward, in the second training round, the field-of-view was activated as shown in Figure 3 such that the participants could get used to the navigation and task with

the limited field-of-view they would have during the experimental trials. During these two training rounds, the AI agent did not execute any actions without being given a command nor send any messages to the human. Once the familiarization was completed, participants continued with the experimental trials. All of the three experimental trials have the same limited field-of-view constraints.

After the trials, we asked the participants to complete a questionnaire to indicate their agreement – on a 5-point Likert scale – with statements about the AI agent’s helpfulness, their overall teamwork, and how cooperative they perceived the agent to be. In addition, we asked participants to rate how intelligent, autonomous, and responsible they perceived the AI agent to be. Our items are mostly based on the ones used in Schermerhorn and Scheutz [63] but were translated into German. All of them are shown in Table 3 (1-9). All questionnaire data were collected with Soscisurvey (version 3.1.06) [24] and the detailed list of questions can be found in Table 3.

In addition, we showed participants replays, i.e., reconstructed videos of their gameplay, for specific examples of the situation types that we defined in Section 3.2.1 and which are presented

in Table 2. We asked participants to watch these replays and answer three questions, i.e., how helpful and appropriate the agent's actions and suggestions were in the shown situation as well as how well it communicated them. The specific items are shown in Table 3 (10-12). The participants were additionally able to write comments or remarks for each situation. Since these replays allow the participants to reflect on their and the agent's behavior and recognize communications or actions that they might have missed during the trials, we let them complete the same questionnaire that they filled out before the replays, again. Participants also reported if they wanted the agent to show more or less initiative and if so, write a comment about which way they would want that. Lastly, they could write any comments or notes they had in general at the end of the experiment.

4.2 Results

We analyzed objective measures of task performance and interactions between the human and the agent, as well as the subjective answers about the participants' perception of their AI teammate.

4.2.1 Task Performance of the Human-AI Team. We evaluated the task reward as defined in Section 3.1.3. In comparison to the fixed autonomy levels, situational autonomy adaptation achieved the highest mean and median reward of 220 (SD = 36). The mean score in the no-autonomy condition was 206 (SD = 34.6), 197.4 (SD = 30.9) in the low-autonomy condition, 203.3 (SD = 30.6) in the moderate-autonomy condition, 210.9 (SD = 36.8) in the high-autonomy condition. These results are visualized in Figure 6 (a). We performed pair-wise comparisons between each combination of conditions to test for significant differences with an independent t-test (with $\alpha = .05$), which revealed that there were no significant differences between the conditions w.r.t team performance score. Since we ran tests for each pair of conditions, we applied a Bonferroni correction for every analysis ($5^{*4/2}$ pairs give a Bonferroni factor of 10).

Additionally, we analyzed the differences in sub-task performance in each condition, i.e., how many books were sorted and folders completed. Most books were sorted in moderate (mean = 25.5 SD = 4.3), adaptive (mean = 25.4 SD = 6.0) and high (mean = 25.1 SD = 5.4) autonomy conditions compared to no (mean = 24.4 SD = 4.5) and low (mean = 23.0 SD = 4.2) autonomy conditions. Participants completed the most folders in the conditions of high-autonomy (mean = 17, SD = 3.21) and adaptive-autonomy (mean = 16.6, SD = 4.23) compared to moderate-autonomy (mean = 14.8, SD = 3.21), low-autonomy (mean = 15.6, SD = 2.51) and no-autonomy (mean = 15.8, SD = 3.63) conditions. While these difference were not statistically significant, we found a significant difference in how many deliveries were accepted between the no-, low- and moderate-autonomy conditions compared to the high- and dynamic-autonomy conditions (all comparisons reveal significant differences with Bonferroni-corrected p-values being < 0.01 , except between no- and high where the p-value is < 0.05). These results are shown in Figure 6 (b).

4.2.2 Interactions with the AI Agent. We investigated how many commands participants gave to the agent and found significant differences between no- and low-autonomy conditions

compared to moderate-, high- and dynamic-autonomy conditions (all significant p-values < 0.001 , Bonferroni-corrected independent t-test). This difference can be seen in Figure 7 (a). Most commands were given for the agent with no autonomy (mean = 30, SD = 5.95). In the low autonomy condition, the agent also received a high amount of commands (mean = 31, SD=3.6). These findings reveal expected interaction patterns, since, by design, these conditions require more commands to achieve cooperative task success. Furthermore, considerably fewer commands were given in the moderate (mean = 19, SD=7), high (mean = 16, SD = 6.4), and adaptive autonomy conditions (mean = 17.5, SD = 5.4).

In addition to the amount of commands given by the human, we also analyzed the amount of messages generated by the AI agent. Figure 7 (b) illustrates these results. Per definition, there are no messages generated by the no autonomy agent. Even though in the low autonomy condition, the agent actively asks the human what to do in case one of the situations described in Table 1 occurs, participants only reacted to 27% of these messages over all 30 trials (598/2181).

In the dynamic and moderate conditions, the agent showed messages to the human to suggest the action it currently considers beneficial and waits for confirmation. It should be noted, that we intentionally implemented this action in a way that the agent does not always recognize the correct label color the human would need. Therefore, in these cases, the agent makes an error and suggests a suboptimal action. Overall, in the moderate autonomy condition, humans accepted 65.6% of the agent suggestions that were presented across all 30 trials (744/1134) and in the dynamic condition they accepted 41.6% (259/622). In the high autonomy condition, the agent only generated messages in case it suggested a change to the human's current actions, e.g., asking the human to put books in the rack. Participants answered those, in only 15% of cases across all 30 trials (22/140).

Figure 7 (c) shows the total amount of interactions between the human and the agent, i.e., the sum of commands and human answers to action suggestions from the agent, i.e., all their communications. We find large significant differences between all combinations (p-value < 0.05 for no-low, low-moderate; p-value < 0.01 for high-dynamic; all other p-values < 0.001). The smallest number of overall interactions occurred in the high autonomy condition (mean = 8.4, SD = 8.9) followed by adaptive (mean = 13, SD = 6.5), moderate (mean = 22.35, SD= 6.6), low (mean = 25.9, SD = 8.8), and no (mean = 29.7, SD=5.9) autonomy conditions.

For those cases where the high autonomy agent handed over labels of the wrong color, it is important to note that only 3 participants placed those labels on the trash pile (10 labels over all trials). In all other cases, participants either counteracted the agent's action to provide the wrong label with a command or adapted their own strategy to make use of the offered label.

4.2.3 Subjective Perception of the AI Agent. The participants answered the questionnaire about their subjective perception of their teammate twice, once before watching the replays and once afterward. There were no significant deviations in the subjective answers before and after watching the replays, except for the statements "Agent was cooperative" and "Agent was capable" in the no autonomy condition. In these two cases, subjects on average

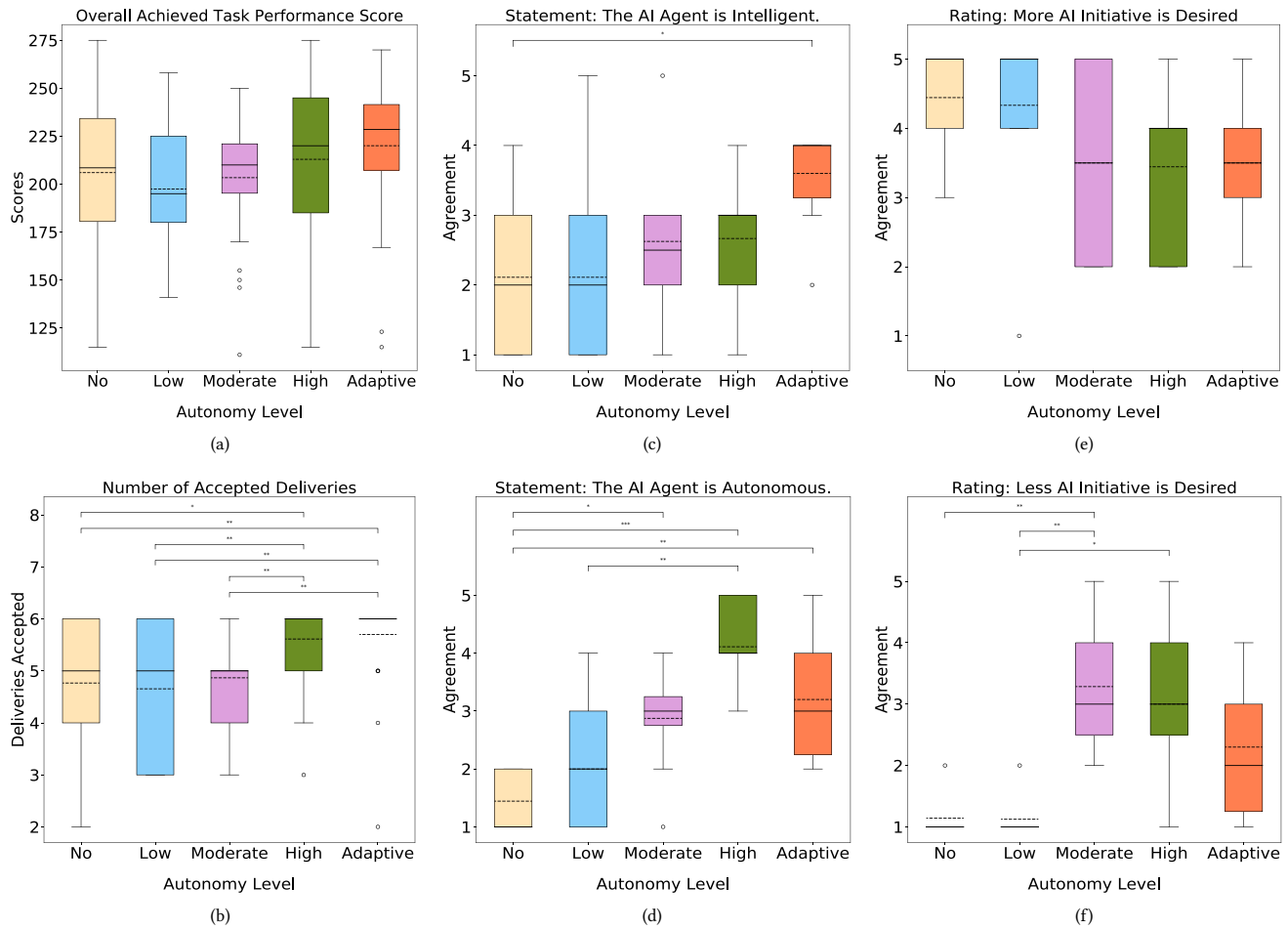


Figure 6: Overview of empirical results comparing fixed and situational adaptation of AI autonomy in our cooperative task within shared workspaces. * indicates a p-value < 0.001, ** stands for p < 0.01 and * shows p < 0.05. All p-values are Bonferroni-corrected taking into account the 10 pairwise comparisons in each subplot. The mean is depicted as a dashed and the median as a solid line. (a) Overall team reward scores, i.e., task performance, in all conditions. (b) Overview of how many deliveries were accepted in each condition. There are significant difference between no-, low- and moderate autonomy conditions compared to the high- and dynamic autonomy conditions. (c) Results showing agreement to the statement “The AI agent is intelligent.” in each condition. (d) Results showing agreement to the statement “The AI agent is autonomous.” in each condition. (e) Results for question about if more initiative is desired from the AI agent in each condition. (f) Results for question about if less initiative is desired from the AI agent in each condition.**

lowered their rating after watching the replays (average rating “cooperative” before replay = 4.3 vs. after replay = 3.0; average rating “capable” before replay = 4.5 vs. after replay = 3.7). Here, we only report the detailed results after the replays, i.e., after the participants had the chance to reflect on their own and the agent’s behavior. The results for all 9 questionnaire items over the five autonomy conditions are presented with their means and standard deviations in Table 3 (1-9). We discuss the results of pair-wise comparisons of independent t-tests (Bonferroni-corrected for each item, like before).

The results show that even though the agent is perceived as most autonomous in the high autonomy condition, it is perceived

as most intelligent when it adapts its autonomy to the current situation. These results are also visualized in Figure 6 (c) and (d), with a significant difference between no and adaptive for the perception of intelligence (p-value = 0.019) and for the perception of agent autonomy between no and moderate (p-value= 0.018). The differences in perceived agent autonomy are also significant between no and both high and adaptive autonomy conditions, as well as between low and high (all p-values < 0.01). Furthermore, there are significant differences between the low condition compared to the high and adaptive ones (both p-values < 0.001) about the agent making its own decisions (item 3 in Table 3). Additionally, there were significant differences between the no autonomy agent

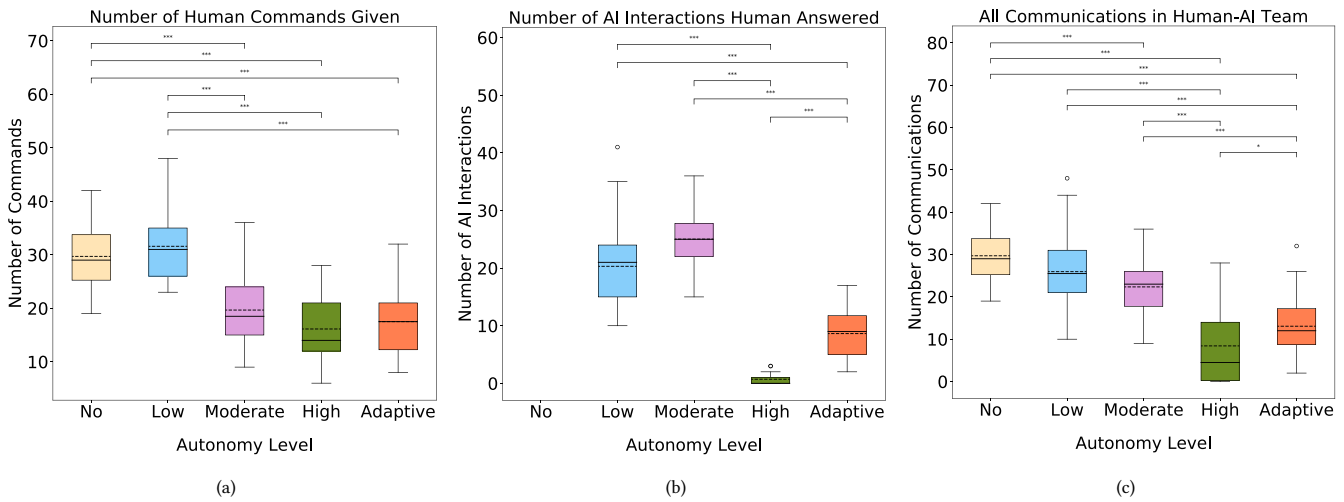


Figure 7: Overview of our empirical results comparing the type of interactions that occurred when comparing fixed and situational adaptation of AI autonomy in our cooperative task within shared workspaces. * indicates a Bonferroni-corrected p-value < 0.001, ** indicates $p < 0.01$ and * shows $p < 0.05$. Mean is depicted as dashed and Median as solid line (a) Overall number of commands given in each condition. (b) Overall number of question from the AI agent that the human reacted to, i.e., answered the agent by e.g., confirming a question like if the agent should bring the human a blue label. (c) Overall interactions within the human-AI team. This includes all communications, i.e., all commands the human gave directly to the agent as well as all answered requests or questions from the AI agent to the human.**

compared to high and adaptive agents (both p-values < 0.05) for perceiving the agent like a team member (item 5 in Table 3) and finally between moderate and adaptive (p-value = 0.041) for the agent’s contribution towards the team goal (item 6 in Table 3).

At the end of the experiment, we asked our participants to indicate whether they would have wanted the agent in their condition to exhibit more or less initiative on a 5-point Likert scale and with free text comments. As shown in Figure 6 (e) in the no- and low-autonomy conditions, more autonomy was clearly desired. The written comments pointed out that participants wished the agent to be more proactive (e.g., sort books, proactively bring labels or tend to the doorbell itself) instead of only waiting for commands. They also remarked that the communication in the low autonomy condition was missed frequently because they were focused on their own task, and they wished for a change in the communication style (but did not specify in what regard).

Figure 6 (f) visualizes if participants wished for less initiative in the different conditions. In the comments for the moderate autonomy condition, the participants pointed out that they would prefer the agent to execute some tasks, such as sorting books or answering the doorbell, without always asking for their confirmation. In contrast, in the high autonomy condition most participants wished for higher agent accuracy, when they noticed it made mistakes and, e.g., autonomously provided wrong labels. However, regarding the confirmation requests in the dynamic and moderate conditions, one participant reported wanting the labels to not just be suggested but be delivered even if they are “wrong”. Another participant wanted the agent to first inform them which label it will bring over, such that the participant can then — based on that label — decide what

document to process next. Lastly, participants wanted even less initiative when the suggestions required any changes to their own plan. For instance, three participants reported the agent asking them if they could place books on the rack such that it can sort them, but they felt like it was interrupting their own workflow. Thus, they did not like it when the agent made such suggestions.

Finally, we examined how participants rated the agent’s actions and communication when they watched the replays. The replays showed a variety of situations of the situation types in Table 1, that occurred during each participant’s own trials. Note that not all situations occurred equally for every participant, so we have varying numbers of answers for each participant. We find significant differences between no and moderate conditions and high and adaptive conditions for how helpful the agent was rated (item 10 in Table 3) and how well it communicated its actions or suggestions, i.e., item 12 in Table 3 (all p-values < 0.001). Ratings for how appropriate the agent’s action or suggestions were (item 11 in Table 3) show significance between the low and adaptive condition (p-value = 0.00002) and between no and all other conditions (all p-values < 0.001). Additionally, there are significant differences between the perception of the communication and helpfulness between low compared to moderate, high and adaptive (all p-values < 0.001). All mean values (and SDs) of the ratings for all participants in each condition are shown in Table 3 (10–12).

5 DISCUSSION

With the current rapid progress in the field of AI, it is important to not only focus on its technical development, but to also ensure that AI systems are designed to actually be helpful for human

Item/Autonomy Level		No	Low	Mod.	High	Adaptive
1	The agent was helpful.	3.5 (1.3)	4.1 (1.0)	4.5 (0.5)	4.4 (0.6)	4.3 (0.4)
2	The agent was capable.	3.6 (0.9)	4.4 (0.6)	4.5 (0.5)	4.2 (0.7)	3.9 (0.8)
3	The agent appeared to make its own decisions.	1.5 (0.6)	1.6 (0.6)	2.8 (1.3)	3.8 (0.9)	4.1 (0.7)
4	The agent was cooperative.	3.0 (1.5)	3.6 (1.1)	3.8 (0.9)	4.1 (0.9)	4.2 (0.6)
5	The agent acted like a member of the team.	3.0 (0.6)	3.8 (0.7)	4.0 (0.7)	4.3 (0.9)	4.1 (0.5)
6	The agent contributed as much as me to achieving the goal.	2.6 (0.9)	3.2 (1.2)	2.2 (1.2)	3.5 (1.1)	4.0 (0.7)
7	The agent is intelligent.	2.1 (0.9)	2.1 (1.2)	2.6 (1.1)	2.6 (0.9)	3.6 (0.6)
8	The agent is autonomous.	1.4 (0.4)	2.0 (1.0)	2.8 (0.9)	4.1 (0.7)	3.2 (0.9)
9	The agent is responsible.	2.4 (1.2)	3.2 (1.4)	3.2 (1.0)	3.2 (0.8)	3.4 (0.6)
10	How helpful was the agent's action?	2.4 (1.2)	2.9 (1.5)	3.6 (1.3)	3.7 (1.3)	3.8 (1.2)
11	How appropriate was the agent's action or suggestion?	2.8 (1.3)	3.3 (1.3)	3.8 (1.2)	3.7 (1.2)	3.8 (1.1)
12	How well did the agent communicate its action?	1.8 (1.0)	2.5 (1.4)	4.1 (1.1)	3.8 (1.2)	4.0 (1.0)

Table 3: All items used in our questionnaire which all participants in each condition answered. The first five items are based on [63], and items 6-9 the remaining ones were added by us. Furthermore, the items that were used during the replays are shown in 10-12. In the table, we report the mean values of agreement the participants indicated and the standard deviation in brackets.

users. Thus, it is necessary that the development process is human-centered [66, 79, 80]. Especially in shared workspaces, where AI systems are not just used as tools but rather need to function as a teammate to the human, it is crucial to design them to be collaborative in order to improve task performance. Thus, in order to test the effectiveness of human-AI teams, we focus on a setting that, by design, requires the teammates to work together to complete the task successfully instead of settings in which the human or the agent could complete the overall task on their own. An essential factor to consider in such human-AI interactions is the AI teammate's autonomy level [50, 51]. In some cases, e.g., in tasks where full sub-task automation is preferred by users and AI systems do not make any mistakes, it will be beneficial to constantly have high AI autonomy [50, 51]. In other cases, when humans want to remain in control and there are high stakes or an AI system's capabilities are limited, it will be better if AI systems remain only on low(er) autonomy levels [2, 33]. However, many real-world applications of AI systems will most likely contain situations with aspects of both cases, i.e., there will be low-stakes situations that are well within the AI agent's capabilities and there will be high-stakes situations that the agent cannot deal with. For such applications, we consider the ability of an AI agent to dynamically adapt its own degree of autonomy between sub-tasks and situations essential. This is particularly important in settings in which humans and AI agents have complementary skills, and thus only perform successfully when they have more autonomy in their area of expertise and less in other parts of the task. In the following, we discuss the main findings from our empirical evaluation, in which we examined effects of four fixed AI autonomy levels and automatic adaptation between them in a cooperative shared workspace task.

5.1 Task Performance of Human-AI Teams in Shared Workspaces

The evaluation of the task performance in our cooperative shared workspace setting showed that the agent that could adapt its autonomy in different situations outperformed the agents in the fixed autonomy levels on average (RQ 1). Even if the differences in the overall performance score were not significant, there were, e.g., clear benefits in the number of accepted deliveries (see Figure 6 (b)). Human behavior clearly changed between different conditions depending on how much they had to control the agent, or not (see Figure 7). A high amount of agent autonomy was generally advantageous in our setting, which aligns well with the work of McNeese et al. [50, 51] who describe higher autonomy as necessary for agents to be helpful within HATs during complex tasks. However, in situations where the agent is uncertain or would interfere with the human's plan, our results show that it is better to reduce autonomy. When the agent made an error in the high autonomy condition, e.g., provided wrong labels, we see that humans tend to compensate by adapting their own behavior such that these errors do not necessarily reflect directly in the overall task performance.

5.2 Human Perception of AI Teammate

Our experiment provides valuable insights on how the human's perception of their AI teammate and the interaction with it is impacted by situational autonomy adaptation as opposed to fixed autonomy levels (RQ 2). While the agent in the high autonomy condition is rated significantly more autonomous than the agents in all other conditions (see Figure 6 (d)), the agent with situationally adaptive autonomy was rated clearly more intelligent compared to all other agents (see Figure 6 (c)). Hence, when agents appropriately adjust their autonomy level depending on the situation, and in particular also decrease it when it is necessary, they were perceived as more intelligent.

We asked the participants whether they wished for more or less initiative of the agent in their respective condition. They overall wanted fewer changes to the way the agent showed initiative in the adaptive autonomy condition. In the no and low autonomy conditions, participants wished for features of the higher autonomy condition, e.g., that the agent should tend to the doorbell on its own or proactively provide labels. While the agent in the high autonomy condition was generally perceived as more helpful and cooperative than in no or low autonomy, participants reported that its errors, i.e., offering wrong labels, influenced its perceived helpfulness for them. For instance, one participant explicitly stated “How helpful the agent was depended on which label it gave me. Sometimes it was right, and sometimes it was wrong”⁴. In such instances, the participants wished for less autonomy of the agent. When participants described which changes to the agent’s exhibited initiative they would want, interestingly, in the high and adaptive autonomy condition, i.e., when the agent already was much more proactive and autonomous, they tended to wish for even more intelligent behavior. In particular, participants wanted more anticipation from the agent and more team planning. Additionally, they wished it would notice general patterns in their behavior and learn to adapt accordingly. One participant stated this as: “Ideally, it should have understood and adapted to my pattern. For example, I always tried to empty the books first. It could understand that it should directly sort the books before doing anything else.”

Even though participants in the high- and dynamic-autonomy conditions wanted more “intelligent” behavior, interestingly, they were not receptive to the agent’s suggestions about changing their own actions. Many participants remarked that the agent should refrain from such suggestions with, e.g., one participant commenting: “The questions were mostly appropriate but its request (for me to change what I’m doing) were inappropriate and going against my own plans”⁴. Hence, while occasionally these suggestions were seen as reminders and were appreciated, generally, our participants would rather like the agent to adjust to their behaviors, or they felt the agent’s understanding of their behavior would need to be better. This is illustrated by a quote when the agent asked a human to answer the doorbell, since the human was closer to the doorbell than the agent, and the human was irritated by the agent’s request, stating “I would have accepted the delivery anyway as I was on my way there to place a folder in the shelf”⁴.

5.3 Limitations and Future Directions

Participants had to switch between the arrow keys and the mouse for replying to the agent’s suggestions or to issue commands. This led to some subjects reporting that they were more hesitant and less willing to answer the agent’s requests or interact with it.

Furthermore, since the agent’s communication messages were presented as visual cues on the screen (text and buttons), participants reported that they sometimes missed these messages because they were too focused on their own parts of the task. Cognitive load might have been too high in some of these cases, which is often found in complex tasks [e.g. 11]. Since employing AI agents as teammates is probably most useful in cognitively challenging

scenarios, our proposed task setup offers a good approximation to further investigate such effects.

This point is also underpinned by our participants wishing for more active planning of the agent, in particular when they were busy with another task. For instance, they wanted it to “recognize their patterns” in order to adjust its behavior accordingly, instead of interrupting them when they are busy and propose a change of plans. Moreover, participants wanted the agent to always tend to the doorbell as they liked when they did not have to think about that sub-task themselves. To possibly alleviate cognitive strain in such situations, participants also suggested the agent’s communication to be implemented in a multimodal fashion, e.g., by the using auditory signals. Overall, an important aspect for further investigation is to investigate HAT dynamics in situations with elevated cognitive strain to explore if humans continually perceive AI cooperation as valuable when deeply immersed in their tasks.

We observed that participants preferred the higher autonomy levels, which also had clear effects on the human’s behavior (e.g., fewer commands and interactions). However, the overall performance did not increase too much compared to lower autonomy levels. The task was challenging, but seemingly not enough to reveal big performance effects of AI support because the humans compensated for the agent’s errors and could achieve a high performance by basically manually controlling the agent.

Furthermore, humans compensated for the agent’s errors, e.g., when it handed over a wrong label, by either preparing the corresponding document or by throwing away the wrong label and requesting a new one. These compensations led to small time delays, but did not strongly influence overall task success, as only sometimes fewer folders could be processed overall. Theoretically, participants could have made errors themselves as well, such as throwing away labels that they could have used for their prepared documents. However, such cases did not occur in our data. Additionally, participants occasionally distributed the sub-tasks somewhat inefficiently. In such cases, the agent in moderate, high and dynamic autonomy conditions suggested a different task distribution. For example, if a label was already provided by the agent, but the human partner sorted books instead of completing the document processing, the agent suggested that the human should rather place the books on its rack. This way the agent could sort the books and the human could complete the part of the task that only they could do. However, we observed that humans often ignored these suggestions, and felt interrupted by them. There were no major errors either humans or the agent could commit, except for not tending to the doorbell, which happened in some cases, as also illustrated in Figure 6 (b). An interesting line for future work is therefore to consider other settings, in which more sources of major and minor errors on both sides are included, especially since with more points in which failures can occur dynamically adjusting the agent’s autonomy might be even more beneficial. This way, on the one hand, the agent could decrease its autonomy when it is likely to make a mistake and, on the other hand, increase its autonomy when it prevents (human) errors, possibly leading to overall better performance outcomes.

In general, our experimental findings provide valuable insights for human-AI collaboration in similar cooperative shared workspaces, i.e., settings in which humans can not complete all sub-tasks

⁴translated from German

by themselves, are willing to give some (autonomous) sub-task control to AI agents and ToMMs are necessary and beneficial to infer a partner's goals and state of knowledge. In particular, when HATs interact in settings in which they face situations such as the ones described in our paper, e.g., when higher priority sub-tasks can arise, humans and agents can make errors, and an agent may make alternative suggestions about the task distribution, our empirical insights can be transferable to implement situational autonomy adaptation for seamless human-AI cooperation. However, the situation types and autonomy levels that we presented in Table 1 are not necessarily exhaustive, and additional ones could be explored in the future. Integrating sophisticated ToMMs in future AI agents could allow them to better adjust to human behavior and, e.g., include a certainty estimate instead of fixed rules for autonomy adaptation [5]. Additionally, in the long run, an agent should be able to learn from experience and adapt to the preferences of its human teammate and discover what autonomy level works best in different situations.

6 CONCLUSION

In this paper, we propose a cooperative task in a simulated shared workspace, in which we investigated how different autonomy levels and their adaptation influence task performance and human perception of the AI agent. We derived adaptation rules for AI autonomy in cooperative shared workspace settings from previous work and from empirical results of a pilot study. In particular, we identify and propose five criteria for implementing the switch of autonomy levels. We evaluated the effects of fixed levels of AI autonomy and situation-dependent autonomy adaptation in a user study. We find that overall team performance was best in the condition where humans collaborated with the agent that adjusted its autonomy based on the situation. Additionally, our results show that not only is the agent with adaptive autonomy perceived as cooperative and helpful, but it was also rated highest in terms of perceived intelligence. In summary, we show that automatically adapting the AI agent's autonomy level depending on the current situation has positive effects on human-AI collaboration in shared workspace settings.

ACKNOWLEDGMENTS

The authors would like to thank all subjects for their participation in the experiments. This work was funded by German Federal Ministry of Education and Research (project IKIDA, 01IS20045).

REFERENCES

- [1] Hussein A Abbass. 2019. Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust. *Cognitive Computation* 11, 2 (2019), 159–171.
- [2] Alper Alan, Enrico Costanza, Joel Fischer, Sarvapali Ramchurn, Tom Rodden, and Nicholas R Jennings. 2014. A Field Study of Human-Agent Interaction for Electricity Tariff Switching. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014)*. International Foundation for Autonomous Agents and Multiagent Systems, France, 965–972.
- [3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fournay, Besmira Nushi, Penny Collisson, Shamsi Iqbal Jina Suh, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)* (Glasgow, Scotland Uk) (CHI '19). ACM, New York, NY, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300233>
- [4] Nantheera Anantrasrichai and David Bull. 2022. Artificial Intelligence in the Creative Industries: A Review. *Artificial intelligence review* 55, 1 (2022), 1–68.
- [5] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. 2017. Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing. *Nature Human Behaviour* 1 (2017), 1–10. Issue 64.
- [6] Matthew Ball and Vic Callaghan. 2012. Explorations of Autonomy: an Investigation of Adjustable Autonomy in Intelligent Environments. In *2012 Eighth International Conference on Intelligent Environments*. IEEE, IEEE Computer Society, USA, 114–121.
- [7] Jenay M Beer, Arthur D Fisk, and Wendy A Rogers. 2014. Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. *Journal of Human-Robot Interaction* 3, 2 (2014), 74.
- [8] RICHARD BELLMAN. 1957. A Markovian Decision Process. *Journal of Mathematics and Mechanics* 6, 5 (1957), 679–684. <http://www.jstor.org/stable/24900506>
- [9] Justin Bishop, Jaylen Burgess, Cooper Ramos, Jade B Driggs, Tom Williams, Chad C Tossell, Elizabeth Phillips, Tyler H Shaw, and Ewart J de Visser. 2020. CHAOPT: A Testbed for Evaluating Human-Autonomy Team Collaboration Using the Video Game Overcooked! 2. In *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, IEEE, USA, 1–6.
- [10] Jeffrey M Bradshaw, Paul J Feltoch, Hyuckchul Jung, Shrinivas Kulkarni, William Taysom, and Andrzej Uszok. 2004. Dimensions of Adjustable Autonomy and Mixed-Initiative Interaction. In *Agents and Computational Autonomy: Potential, Risks, and Solutions 1*. Springer, Springer, Berlin, Germany, 17–39.
- [11] Yannick Brand and Axel Schulte. 2021. Workload-Adaptive and Task-Specific Support for Cockpit Crews: Design and Evaluation of an Adaptive Associate System. *Human-Intelligent Systems Integration* 3 (2021), 187–199.
- [12] David J Bruemmer, JL Marble, Matthew O Anderson, MD McKay, and DD Dudenhoeffer. 2002. Dynamic-Autonomy for Remote Robotic Sensor Deployment.
- [13] Andres Campero, Michelle Vaccaro, Jaeyoon Song, Haoran Wen, Abdullah Al-maatouq, and Thomas W. Malone. 2022. A Test for Evaluating Performance in Human-Computer Systems. [arXiv:2206.12390 \[cs.HC\]](https://arxiv.org/abs/2206.12390)
- [14] Stuart K Card, Thomas P Moran, and Allen Newell. 2018. *The Psychology of Human-Computer Interaction*. Crc Press, Boca Raton.
- [15] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-AI coordination. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 465, 12 pages.
- [16] Cristiano Castelfranchi. 2000. Founding Agents' Autonomy on Dependence Theory. In *ECAI*, Vol. 1. IOS Press, NLD, 353–357.
- [17] Mustafa Mert Çelikok, Tomi Peltola, Pedram Daei, and Samuel Kaski. 2019. Interactive AI with a Theory of Mind.
- [18] Eftychios G Christoforou, Andreas S Panayides, Sotiris Avgousti, Panicos Moursas, and Constantinos S Pattichis. 2020. An Overview of Assistive Robotics and Technologies for Elderly Care. In *XV Mediterranean Conference on Medical and Biological Engineering and Computing—MEDICON 2019: Proceedings of MEDICON 2019, September 26–28, 2019, Coimbra, Portugal*. Springer, Springer International Publishing, Cham, 971–976.
- [19] Ylona Chun Tse, Melanie Birks, and Karen Francis. 2019. Grounded Theory Research: A Design Framework for Novice Researchers. *SAGE open medicine* 7 (2019), 2050312118822927.
- [20] Nancy Cooke, Mustafa Demir, and Lixiao Huang. 2020. A Framework for Human-Autonomy Team Research. In *Engineering Psychology and Cognitive Ergonomics, Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*. Springer, Springer International Publishing, Cham, 134–146.
- [21] J.W. Crandall and M.A. Goodrich. 2001. Experiments in adjustable autonomy. In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236)*, Vol. 3. IEEE, USA, 1624–1629 vol.3. <https://doi.org/10.1109/ICSMC.2001.973517>
- [22] Mustafa Demir, Nathan J McNeese, and Nancy J Cooke. 2017. Team Situation Awareness within the Context of Human-Autonomy Teaming. *Cognitive Systems Research* 46 (2017), 3–12.
- [23] Sandra Devin and Rachid Alami. 2016. An Implemented Theory of Mind to Improve Human-Robot Shared Plans Execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, IEEE, USA, 319–326.
- [24] D.J. Leiner. 2018. SosciSurvey. <https://www.socisurvey.de/en/index>
- [25] Gregory Dorais, R Peter Bonasso, David Kortenkamp, Barney Pell, and Debra Schreckenghost. 1998. Adjustable Autonomy for Human-Centered Autonomous Systems on Mars.
- [26] Mica R Endsley. 2017. From Here to Autonomy: Lessons Learned from Human-Automation Research. *Human Factors* 59, 1 (2017), 5–27.
- [27] Mica R Endsley. 2018. Automation and Situation Awareness. In *Automation and human performance: Theory and applications*. CRC Press, USA, 163–181.
- [28] Maurizio Faccio, Irene Granata, Alberto Menini, Mattia Milanese, Chiara Rossato, Matteo Bottin, Riccardo Minto, Patrik Pluchino, Luciano Gamberini, Giovanni Boschetti, et al. 2023. Human Factors in Cobot Era: A Review of Modern Production Systems Features. *Journal of Intelligent Manufacturing* 34, 1 (2023), 85–106.

- [29] Michelangelo Fiore, Aurélie Clodic, and Rachid Alami. 2016. On Planning and Task achievement Modalities for Human-Robot Collaboration. In *Experimental Robotics: The 14th International Symposium on Experimental Robotics*. Springer, Springer International Publishing, Cham, 293–306.
- [30] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM New York, NY, USA, 1–12.
- [31] Michael A Goodrich, Timothy W McLain, Jeffrey D Anderson, Jisang Sun, and Jacob W Crandall. 2007. Managing Autonomy in Robot Teams: Observations from Four Experiments. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, 25–32. <https://doi.org/10.1145/1228716.1228721>
- [32] Xiao Guo, Zhenjiang Shen, Yajing Zhang, and Teng Wu. 2019. Review on the Application of Artificial Intelligence in Smart Homes. *Smart Cities* 2, 3 (2019), 402–420.
- [33] Allyson I Hauptman, Beau G Schelble, Nathan J McNeese, and Kapil Chalil Madathil. 2023. Adapt and Overcome: Perceptions of Adaptive Autonomous Agents for Human-AI Teaming. *Computers in Human Behavior* 138 (2023), 107451.
- [34] Ziyao He, Yunpeng Song, Shurui Zhou, and Zhongmin Cai. 2023. Interaction of Thoughts: Towards Mediating Task Assignment in Human-AI Cooperation with a Capability-Aware Shared Mental Model. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18.
- [35] Felix Heilemann, Sebastian Lindner, and Axel Schulte. 2021. Experimental Evaluation of Tasking and Teaming Design Patterns for Human Delegation of Unmanned Vehicles. *Human-Intelligent Systems Integration* 3 (2021), 223–240.
- [36] Andreas Hepp. 2020. Artificial Companions, Social Bots and Work Bots: Communicative Robots as Research Objects of Media and Communication Studies. *Media, Culture & Society* 42, 7-8 (2020), 1410–1426.
- [37] Laura M Hiatt, Anthony M Harrison, and J Gregory Trafton. 2011. Accommodating Human Variability in Human-Robot Teams through Theory of Mind. In *Twenty-Second International Joint Conference on Artificial Intelligence* (Barcelona, Catalonia, Spain) (*IJCAI'11*). AAAI Press, USA, 2066–2071.
- [38] Kenneth Holstein and Vincent Alevan. 2021. Designing for Human-AI Complementarity in K-12 Education. arXiv:2104.01266 [cs.HC] <https://arxiv.org/abs/2104.01266>
- [39] Toshiyuki Inagaki et al. 2003. Adaptive Automation: Sharing and Trading of Control. *Handbook of cognitive task design* 8 (2003), 147–169.
- [40] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, Libuše Hannah Vepřek, and Gabrielle Quinn. 2022. Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making. *ACM Transactions on Computer-Human Interaction* 30, 5 (2022), 1–29. <https://doi.org/10.1145/3534561>
- [41] Jinglu Jiang, Alexander J Karran, Constantinos K Coursaris, Pierre-Majorique Léger, and Joerg Beringer. 2022. A Situation Awareness Perspective on Human-AI Interaction: Tensions and Opportunities. *International Journal of Human-Computer Interaction* 39, 9 (2022), 1–18.
- [42] David Kieras. 2004. GOMS Models for Task Analysis. *The Handbook of Task Analysis for Human-Computer Interaction* 1 (2004), 83–116.
- [43] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18.
- [44] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. 2023. Trusting Artificial Agents: Communication Trumps Performance. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 299–306.
- [45] Bennie Lewis, Bulent Tastan, and Gita Sukthankar. 2013. An Adjustable Autonomy Paradigm for Adapting to Expert-Novice Differences. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, IEEE, USA, 1656–1662.
- [46] Brian Y Lim and Anind K Dey. 2009. Assessing Demand for Intelligibility in Context-Aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*. Association for Computing Machinery, New York, NY, USA, 195–204.
- [47] Lanny Lin and Michael A Goodrich. 2015. Sliding Autonomy for UAV Path-Planning: Adding New Dimensions to Autonomy Management. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1615–1624.
- [48] Rachel Macrorie, Simon Marvin, and Aidan While. 2021. Robotics and Automation in the City: A Research Agenda. *Urban Geography* 42, 2 (2021), 197–217.
- [49] Alexander Maedche, Christine Legner, Alexander Benlian, Benedikt Berger, Henner Gimpel, Thomas Hess, Oliver Hinz, Stefan Morana, and Matthias Söllner. 2019. AI-based Digital Assistants: Opportunities, Threats, and Research Perspectives. *Business & Information Systems Engineering* 61 (2019), 535–544.
- [50] Nathan J McNeese, Mustafa Demir, Nancy J Cooke, and Christopher Myers. 2018. Teaming with a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human factors* 60, 2 (2018), 262–273.
- [51] Nathan J McNeese, Beau G Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. 2021. Who/What is my Teammate? Team Composition Considerations in Human-AI Teaming. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 288–299.
- [52] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, NY, USA, 1–34.
- [53] Vera Zaychik Moffitt, Jerry L Franke, and Meghann Lomas. 2006. Mixed-Initiative Adjustable Autonomy in Multi-Vehicle Operations. In *Proceedings of AU/VS, Orlando, Florida*. Association for Unmanned Vehicle Systems International (AU/VS), USA.
- [54] Salama A Mostafa, Mohd Sharifuddin Ahmad, and Aida Mustapha. 2019. Adjustable Autonomy: A Systematic Literature Review. *Artificial Intelligence Review* 51, 2 (2019), 149–186.
- [55] Geoff Musicik, Rui Zhang, Nathan J McNeese, Guo Freeman, and Anurata Prabha Hridi. 2021. Leveling up Teamwork in Esports: Understanding Team Cognition in a Dynamic Virtual Environment. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–30.
- [56] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human factors* 64, 5 (2022), 904–938.
- [57] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 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 (2000), 286–297.
- [58] Marc Pinski, Martin Adam, and Alexander Benlian. 2023. AI Knowledge: Improving AI Delegation through Human Enablement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–17.
- [59] Md Khurram Monir Rabby, Ali Karimoddini, Mubbashar Altaf Khan, and Steven Jiang. 2022. A Learning-Based Adjustable Autonomy Framework for Human-Robot Collaboration. *IEEE Transactions on Industrial Informatics* 18, 9 (2022), 6171–6180.
- [60] Anjana Ramkumar, Pieter Jan Stappers, Wiro J Niessen, Sonja Adebahr, Tanja Schimek-Jasch, Ursula Nestle, and Yu Song. 2017. Using GOMS and NASA-TLX to Evaluate Human-Computer Interaction Process in Interactive Segmentation. *International Journal of Human-Computer Interaction* 33, 2 (2017), 123–134.
- [61] Thomas M Roehr and Yuping Shi. 2010. Using a Self-Confidence Measure for a System-Initiated Switch Between Autonomy Modes. In *Proceedings of the 10th international symposium on artificial intelligence, robotics and automation in space, Sapporo, Japan*. ESA, France, 507–514.
- [62] David Rother, Thomas Weisswange, and Jan Peters. 2023. Disentangling Interaction using Maximum Entropy Reinforcement Learning in Multi-Agent Systems. In *European Conference on Artificial Intelligence*. IOS Press, NLD, 1994–2001.
- [63] Paul Schermerhorn and Matthias Scheutz. 2009. Dynamic Robot Autonomy: Investigating the Effects of Robot Decision-Making in a Human-Robot Team Task. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*. Association for Computing Machinery, New York, NY, USA, 63–70.
- [64] Thomas B Sheridan. 1992. *Telerobotics, Automation, and Human Supervisory Control*. MIT press, MA, USA.
- [65] Thomas B Sheridan and William L Verplank. 1978. *Human and Computer Control of Undersea Teleoperators*. Technical Report. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.
- [66] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124.
- [67] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian Modeling of Human-AI Complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119.
- [68] Anselm Strauss and Juliet Corbin. 1994. *Grounded Theory Methodology: An Overview*. Sage Publications, Inc, USA, 273–285.
- [69] S Shyam Sundar. 2020. Rise of Machine Agency: A Framework for Studying the Psychology of Human-AI Interaction (HAI). *Journal of Computer-Mediated Communication* 25, 1 (2020), 74–88.
- [70] K Suzanne Barber, Anuj Goel, and Cheryl E Martin. 2000. Dynamic Adaptive Autonomy in Multi-Agent Systems. *Journal of Experimental & Theoretical Artificial Intelligence* 12, 2 (2000), 129–147.
- [71] Milind Tambe, David V Pynadath, Nicholas Chauvat, Abhimanyu Das, and Gal A Kaminka. 2000. Adaptive Agent Integration Architectures for Heterogeneous Team Members. In *Proceedings Fourth International Conference on Multiagent Systems*. IEEE, IEEE, USA, 301–308.
- [72] George Terzopoulos and Maya Satratzemi. 2020. Voice Assistants and Smart Speakers in Everyday Life and in Education. *Informatics in Education* 19, 3 (2020), 473–490.

- [73] Basil Wahn and Alan Kingstone. 2021. Humans Share Task Load with a Computer Partner if (They Believe that) it Acts Human-Like. *Acta Psychologica* 212 (2021), 103205.
- [74] James C Walliser, Ewart J de Visser, Eva Wiese, and Tyler H Shaw. 2019. Team Structure and Team Building Improve Human–Machine Teaming with Autonomous Agents. *Journal of Cognitive Engineering and Decision Making* 13, 4 (2019), 258–278.
- [75] Justin D Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection not Required? Human-AI Partnerships in Code Translation. In *26th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, USA, 402–412. <https://doi.org/10.1145/3397481.3450656>
- [76] Henry M Wellman. 1992. *The Child's Theory of Mind*. The MIT Press, MA, USA.
- [77] Carolin Wienrich and Marc Erich Latoschik. 2021. Extended Artificial Intelligence: New Prospects of Human-AI Interaction Research. *Frontiers in Virtual Reality* 2 (2021), 686783.
- [78] Sarah A Wu, Rose E Wang, James A Evans, Joshua B Tenenbaum, David C Parkes, and Max Kleiman-Weiner. 2021. Too Many Cooks: Bayesian Inference for Coordinating Multi-Agent Collaboration. *Topics in Cognitive Science* 13, 2 (2021), 414–432.
- [79] Wei Xu. 2019. Toward Human-Centered AI: A Perspective from Human-Computer Interaction. *Interactions* 26, 4 (2019), 42–46.
- [80] Wei Xu, Marvin J Dainoff, Liezhong Ge, and Zaifeng Gao. 2023. Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. *International Journal of Human–Computer Interaction* 39, 3 (2023), 494–518.
- [81] Kimitoshi Yamazaki, Ryohei Ueda, Shunichi Nozawa, Mitsuharu Kojima, Kei Okada, Kiyoshi Matsumoto, Masaru Ishikawa, Isao Shimoyama, and Masayuki Inaba. 2012. Home-Assistant Robot for an Aging Society. *Proc. IEEE* 100, 8 (2012), 2429–2441.
- [82] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human" Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.
- [83] Stéphane Zieba, Philippe Polet, Frédéric Vanderhaegen, and Serge Debernard. 2010. Principles of Adjustable Autonomy: A Framework for Resilient Human-Machine Cooperation. *Cognition, Technology & Work* 12, 3 (2010), 193–203.