Designing Human-Agent Collaborations
Commitment, responsiveness, and support

Cila, Nazli

DOI
10.1145/3491102.3517500

Publication date
2022

Document Version
Final published version

Published in
CHI 2022 - Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems

Citation (APA)
Cila, N. (2022). Designing Human-Agent Collaborations: Commitment, responsiveness, and support. In C. Lampe, & S. Barbosa (Eds.), *CHI 2022 - Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* Article 420 (Conference on Human Factors in Computing Systems - Proceedings). Association for Computing Machinery (ACM). https://doi.org/10.1145/3491102.3517500

Designing Human-Agent Collaborations: Commitment, responsiveness, and support

Nazli Cila
Delft University of Technology, Department of Human-Centered Design
n.cila@tudelft.nl

ABSTRACT

With the advancements in AI, agents (i.e., smart products, robots, software agents) are increasingly capable of working closely together with humans in a variety of ways while benefiting from each other. These human-agent collaborations have gained growing attention in the HCI community; however, the field lacks clear guidelines on how to design the agents' behaviors in collaborations. In this paper, the qualities that are relevant for designers to create robust and pleasant human-agent collaborations were investigated. Bratman's Shared Cooperative Activity framework was used to identify the core characteristics of collaborations and survey the most important issues in the design of human-agent collaborations, namely code-of-conduct, task delegation, autonomy and control, intelligibility, common ground, offering help and requesting help. The aim of this work is to add structure to this growing and important facet of HCI research and operationalize the concept of human-agent collaboration with concrete design considerations.

CCS CONCEPTS

• **Human-centered computing**; • **Human computer interaction (HCI)**; • **HCI theory, concepts and models**;

KEYWORDS

human-agent collaboration, Shared Cooperative Activity, autonomous agent, design

ACM Reference Format:
Nazli Cila. 2022. Designing Human-Agent Collaborations: Commitment, responsiveness, and support. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29–May 05, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3491102.3517500

1 INTRODUCTION

The success of Djokovic in Wimbledon 2019 was partially attributed to his smart tennis racket with its accompanying AI algorithm [93]. This smart product helped him to understand his play patterns and how to improve his performance, as well as which strategies to employ against specific opponents. Similar partnerships have long been established with robots, for example in healthcare (e.g. [89]), education (e.g. [13]), manufacturing (e.g. [154]), military (e.g. [38]), among many other fields. AI-powered software agents as well have started to work together with doctors (e.g. [146]), journalists (e.g. [106]), and musicians (e.g. [101]) with varying degrees of success.

This paper is focused on such symbiotic relationships between humans and agents, which are entitled as "human-agent collaboration". The word "agent" is used for referring to artefacts that are equipped with abilities to sense the user and/or environment, autonomously act on the collected data, communicate with users and each other, learn and evolve [29]. This broad definition deliberately aims to encompass smart products, robots, and software agents; and is in line with the "weak notion of agency" described in [160], where an agent is defined as an artefact with autonomy, some form of social ability, reactivity, and proactiveness.

The increasingly intertwined complementary interactions between humans and agents have been an abiding interest in the Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI) communities, which gained even more attention in the recent years with the advancements in AI. These mutually amplifying relationships were presented under the names of Man-Machine Symbiosis [92], Symbiotic Interaction [72], Human-Agent Collectives [74], Human-Computer Integration [49], Cooperative Intelligence [132], Cooperative AI [37], Human-AI Teaming [45], Collaborative control [52], and Mixed-initiative interaction [23]. Although there are differences in the disciplines from which these concepts come (i.e., HCI, HRI, AI, engineering) and how these concepts were unpacked, they all share the notion of the computer as a partner and argue that the future of human-computer interaction lies in here. These works excel in presenting this vision and the opportunities and challenges to get there, but as the HCI community, we still lack clear design patterns to guide future design of agent behaviors in collaborations [12, 124, 152]. Stephanidis and Salvendy [137] named human-technology symbiosis as one of the seven HCI grand challenges, in which the field is urgently invited to find ways to integrate intelligence into everyday life, while also preserving human self-efficacy and control.

To achieve this vision, a deeper investigation on human-agent collaboration, its characteristics, and the considerations that need to be taken when designing collaborations are essential. The aim of this paper is to investigate what is required from humans and agents in terms of collaboration capabilities and support future research and design efforts on human-agent collaboration in HCI by depicting the qualities that are relevant for designing these collaborations. In order to understand the characteristics of human-agent collaborations, this paper will draw from one of the most prominent frameworks to explain human-human collaborations, i.e., Shared Cooperative Activity (SCA) of Michael Bratman [17]. Bratman's framework will be used to cluster the current research on agents and survey the pertinent issues in the design of human-agent collaborations. This synthesis of relevant concepts and findings

CHI '22, April 29–May 05, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3517500

from the HCI and HRI literature under the qualities of human-agent collaborations also allows for a range of design considerations to elicit. In total, there are 11 design considerations that are explained throughout the paper. These considerations are posed in the form of questions to the designers and researchers as a trigger to reflect on the multiple dimensions of human-agent collaborations and envision different ways to attain them. The paper also proposes some promising starting points and directions that could be used for addressing these questions.

Employing Bratman's framework do not suggest that human-agent collaborations should be designed to emulate all attributes of human-human collaborations. Although mimicking interactions among humans as closely as possible has the benefit of helping people to apply already existing expectations and mental schemes to the interaction with agents, it also comes with costs (see [67] for examples). Human-human collaboration and human-agent collaboration are different kind of encounters (see for example agentic "superpowers" under Section 3.1.2 and 3.2.2); however, the paper aims to highlight the basic prerequisites and abilities that are indispensable for agents to eventually enable complex, robust, and pleasant collaborations with humans. That is, although human-agent collaborations will for a long time be (and should be) bound to differ from human-human collaborations in their form, function, and feel, agents need to be equipped with core capabilities to make collaborations with humans similarly as effective and pleasant as collaborations between humans. The intended contribution of this paper is therefore two-fold: to bring structure and clarity to the emerging body of work on human-agent collaboration in HCI and HRI, and to operationalize the concept of human-agent collaboration with concrete design considerations.

In the following section, Bratman's SCA framework will be reviewed and the three core characteristics of human-human collaborations will be underlined: commitment to the joint activity, mutual responsiveness, and commitment to support [17]. Then, the current issues discussed in relation to agent design and development will be organized under these characteristics (Section 3). These are code-of-conduct, task delegation, autonomy and control, intelligibility, common ground, offering help and requesting help. In Section 4, the implications of this analysis for HCI knowledge will be discussed and the intended next steps in this research will be described.

2 HUMAN-HUMAN COLLABORATION

Let us suppose that two flat mates, John and Bob, agree to clean their living room. Each intends to do his share of cleaning and trust that the other will do the same. They will be responsive to each other's plans and actions, e.g., John will finish dusting before Bob will start vacuuming. They will adjust their cleaning tasks so that they will not have resource conflicts, e.g., both needing the sink at the same time. They may also help each other out when needed, e.g., Bob may offer to lift the sofa so that John can vacuum underneath.

Human life is full of collaborations such as this. Collaboration entails that those collaborating are engaged in *joint intentional action* [95]. Joint action is a social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment [131]. Even when they may appear identical, a person's processes are very different in individual and joint actions [34]. Coordinating one's actions with others to achieve a joint outcome, such as lifting the sofa, necessitates an interlocking of individual's behaviors, motor commands, action plans, and perceptions (see [84] for a review of joint action). Therefore, some sort of *shared intention* is necessary [95, 143]. Shared intention specifies the representational systems that enable the planning of joint actions [84].

There are a variety of accounts on shared intention (see [143] for a detailed discussion on individualistic or pluralistic views on shared intention). Shared Cooperative Activity (SCA) of Bratman [17] is one of the most prominent accounts on shared agency. The influence of SCA extends outside philosophy to developmental psychology, artificial intelligence, and robotics [15]. Bratman considers shared intentions as a state of affairs consisting of a set of interrelated individual intentional states. When two people have a shared intention to conduct a joint activity, then each intends that they conduct the joint activity together [17]. That is, John and Bob clean the living room together when their cleaning together is the product of each of them acting with an intention of contributing to their joint activity to cleaning the living room together. Furthermore, each intends that this be brought about by way of co-realizable subplans for their cleaning [15]. This requirement that the subplans "mesh" is designed to rule out cases where John and Bob both intend that they clean the living room, but their individual plans about how to go about it undermine the joint action. John may only want to use expensive environmentally-friendly cleaning agents that someone needs to go and buy from the supermarket, but Bob may want to use the cheap cleaning agents that they already have at home. In this case, the individual subplans are in conflict, and this would prevent John and Bob from cleaning the living room theoretically "together". Furthermore, the meshing subplans condition also rules out certain kinds of coercion, manipulation and deception. For instance, it rules out the case where John steals Bob's favorite t-shirt to mop the floors or where Bob deceives John into thinking that he had not add chemical cleaning agents to the bucket (when he in fact has), even though he knows that John's subplan is that they clean with ecological cleaning agents only. If the two parties' intentions interlock and if their beliefs about interdependence and the common knowledge appropriately coordinates an activity that is being done together, then it is a Shared Cooperative Activity, i.e., collaboration, according to Bratman [15].

With this sketch of the requisite shared intentionality in hand, we will now turn to the characteristics that Bratman identified as central to collaboration (other scholars such as Gilbert [55] and Ludwig [95]) also indirectly mention these characteristics in relation to shared intentionality and full cooperation, respectively). These are: commitment to the joint activity, mutual responsiveness, and commitment to support [17]. While Bratman acknowledges that collaborations can involve large numbers of participating individuals and can take place within a complex institutional framework, his analysis specifically focuses on shared cooperative activities that involve only a pair of participating individuals and are not the activities of complex institutions with structures of authority (i.e., excludes, for example, the activities of a symphony orchestra following its conductor). He entitles these egalitarian small-scale shared activities as "modest sociality" [17], which is also our focus in this paper (Section 3 will investigate the collaborations between a human and an agent; and other collaboration configurations will be briefly presented in Section 4). Below, we will elaborate on each collaboration characteristic.

2.1 Commitment to the joint activity

The first characteristic of collaboration is that the participating individuals each have an appropriate commitment to the joint activity. Ludwig [95] defines full cooperation as "participants in a joint action are committed to rendering appropriate contributions as needed toward their joint end so as to bring it about as efficiently as they can, consistently with the type of action and the generally agreed upon constraints within which they work" (p. 10). Similarly, Gilbert [56] maintains obligations with corresponding entitlements inhere in any joint commitment.

Each individual may participate in a collaboration with different intentions [17]. When John and Bob clean together, John may be primarily concerned with having a clean house, while Bob with finding his lost AirPod while tidying up. In collaborations, each individual will typically have such an intention for some reason or other and these intentions will keep them both working on the cleaning until the living room is clean. When committed to a joint activity, an individual cannot hold intentions that conflict with his/her commitment to the collaborative plan, is required to make a new plan if something goes wrong, and take actions to avoid potential for resource conflicts [61].

Commitment to the joint activity also necessitates the individuals to agree on how they are going to do the joint action and concur on who will take on each of these tasks. Here, the individuals must reason about the match between their capabilities and the requirements of the task/context. They also need to reconcile intentions from private acts with those from group activities [61]. These agreements and the obligations they entail play a central role in the mental state required for collaboration.

2.2 Mutual responsiveness

The second characteristic of collaboration is that each participating individual is responsive to the intentions and actions of the other, knowing that the other party is attempting to be similarly responsive [17]. Jointly acting individuals do not merely act in parallel; each responds to what the other do and plan to do [87]. In our cleaning scenario, Bob may start filling in the bucket for mopping after seeing that John is almost done with vacuuming; or John may dry the leaves of the plants that Bob has been spraying. Being mutually responsive requires the individuals to adjust their behavior as needed in light of information about how the other party is faring in order to ensure effective pursuit of their joint end [95].

Bratman [17] identifies two types of mutual responsiveness. The first one is mutual responsiveness *of intention*. It refers to the situations in which each individual constructs his/her subplans to accomplish the joint task with an eye to meshing with the other's subplans. But then each will go off and play their role with no further interaction. Bratman entitles this noninteractive performance as "prepackaged cooperation" (e.g., John and Bob agree that John will vacuum before going to work in the morning and Bob will mop after work in the evening). For a joint activity to be a collaboration, the individuals must additionally be responsive to the actions of the other. This mutual responsiveness *in action* is a defining characteristic of collaborations.

2.3 Commitment to mutual support

The last characteristic has to do with each agent committing to support the efforts of the other to play his/her role in the joint activity [17]. Willingness to aid is also identified as a feature of fully cooperative behavior [95]. This implies that the participating individuals are ready and willing to provide aid to other in performance of their parts in the joint action and do not interfere with the contributions of the other, insofar as it is compatible with their performing their own parts and the aid offered does not otherwise detract from their joint pursuit of the goal [95]. In our scenario, if John believes that Bob needs help to clean the living room windows, he is prepared to provide such help and that Bob is similarly prepared to support John in his tasks. If John assesses that helping with cleaning the windows would cost an extra hour, he may in the end decide he cannot help Bob out, but he must consider the tradeoff between doing so and not. That is, he needs to weigh his obligation to the joint activity against other commitments in light of his abilities (and also the requirements of morality) and make a decision [61]. Some participants in a collaboration may be willing to incur what would normally be seen as fairly high costs in helping the other; others may be willing to help only if the costs thereby incurred are of a sort that would normally be seen as minimal [17]. Regardless of the substantiality of the help, the commitment to support the other differentiates collaboration from other forms of jointly intentional activity.

2.4 Final notes on Human-Human Collaboration

To synthesize what we have so far, the joint act of cleaning the living room is a collaboration between John and Bob because they both have the intention and commitment to cleaning the living room and trust that the other has the same. They agreed on a plan to follow and have a role division. Each seeks to guide his cleaning behavior with an eye to the behavior of the other, knowing that the other seeks to do likewise. If one of them needs help, the other will assist. So, there is commitment to the joint activity, an agreed-upon role division and action plan, mutual responsiveness in action, and commitment to the success of the other's actions. In short, they have a collaboration.

Bob and John also have a means to *communicate* with each other (gestures and speech in this case) that helps guiding their cleaning activities. As a matter of fact, none of the collaboration characteristics can be attained without communication between the collaborating individuals. Cohen and Levesque [31] argued that the collaborating partners would neither be able to establish a common ground nor interweave subplans to achieve a shared goal without a means to effectively communicate with each other. Tollefsen [144] maintained that joint attention, reading intention-in-action, and verbal communication are mandatory for guiding the actions of collaborating individuals. Clark [35] demonstrated how speech is used to prespecify who will do what and to agree on the specifics of the joint performance; and Grosz [61] discussed the importance of communication in resolving conflicts when carrying out a joint activity.

In the next section, we will adapt the collaboration characteristics of Bratman to human-agent collaborations. Bratman's work is no stranger to the field of computer science. Grosz and her colleagues employed it to design architectures for collaboration-capable, resource-bounded computer agents [62, 63]. Cohen and Levesque [31] proposed a formalism that provides analyses for Bratman's commitment characteristic, specifically focusing on the functional roles played by intentions of collaborative agents. In HRI, Chang et al. [26] and Hoffman and Breazeal [69] built dynamic collaborative frameworks for human-robot teams by adapting SCA. All these works excel in operationalizing Bratman's framework to develop algorithms. Our approach in this paper is different that we use this framework to present in a systematic fashion the qualities that are especially important for effective and pleasant human-agent collaborations. In the next section, these qualities will be discussed in detail.

3 HUMAN-AGENT COLLABORATION

Human-agent collaboration is about humans and agents working closely together in a variety of ways, enhancing human experience. These collaborations require humans and agents to be aware of each other's strengths and limitations, negotiate and align intentions, and support each other. When designed properly, human-agent collaborations can amplify human capacity for reasoning, learning, decision-making and problem solving—all leading to flourishing and empowerment.

Table 1 presents a summary of the collaboration qualities that will be discussed in this section, together with the relevant design considerations and some promising starting points for designers to address and/or answer these. In the following sub-sections, these will be discussed in detail.

3.1 Commitment to the joint activity

As mentioned in Section 2.1, commitment to the joint activity concerns the agreements and obligations related to the collaboration. The collaborators must agree on how they will undertake the joint action and decide who will do what. In relation to human-agent collaboration, we will discuss code of conduct and task delegation respectively, as activities intended for laying the groundwork for collaboration.

3.1.1 Code of conduct.

Design consideration 1: Are the agent's intentions and protocols visible to users? Committing to a collaborative activity requires that the participating parties undertake appropriate actions as needed towards the joint task, while they trust that the other party will do the same [95]. This means that the participants have *task responsibility*—i.e., being responsible for the assigned task—towards each other. Bratman [18] further argues that mutual obligations and entitlements are extremely common in cases of collaboration, which are moral obligations associated with assurance, reliance, promises, and the like. This refers to an additional *moral responsibility* that the parties have towards each other—i.e., being responsible for one's intentions and actions. In order to collaborate with an agent, humans need assurance that the agent bears both types of responsibility.

In relation to autonomous systems, Santoni de Sio and van den Hoven [128] unpacked the concept of *meaningful human control* based on the insights from the literature on free will and moral responsibility. Their philosophical account was originally aimed at autonomous weapon systems, yet their insights have been generalized to IoT devices [147], surgical robots [50] and dual-mode driving systems [102], among others. The authors propose two conditions that should be jointly satisfied in order to achieve meaningful human control over intelligent systems, namely *tracking* and *tracing* [128]. The tracking condition requires a system to be responsive to the relevant human reasons to act. In order to meet the tracking condition, one should identify the relevant human agents and moral reasons at stake in different scenarios with the system, and the level of responsiveness to those moral reasons under different circumstances. Tracing condition requires the presence of one or more human agents in the system's design history or use context who can at the same time understand the capabilities of the system and recognize their own moral responsibility for the system's behavior. The joint satisfaction of these two conditions grants human controllers, designers, programmers, regulators and others a more meaningful kind of control over automated systems, thereby maximizing safety and eliminating unwanted accountability gaps [102].

These may seem as requirements addressed to the creators of autonomous agents; however, van den Hoven [149] argued that the users of autonomous systems also have a so-called "meta-task responsibility". That is, an obligation to check whether the system allows them to do what they ought to do in situ. Such responsibilities are also discussed in relation to robot-assisted surgeries where surgeons and robots collaborate. Fosch-Villaronga et al. [53], for example, maintain that the role of human performance in medical robot surgeries may decrease as robotic technology takes over, but the role of human oversight will increase. A similar trend is being seen in "AI-chaperones" in autonomous driving [83]. These point to an additional *overseeing responsibility* that humans are required to have in human-agent collaborations, where they are required to keep an eye on the agent regarding whether it fulfills its task and moral responsibilities.

Meaningful human control is an important construct because it brings forward the important issues where human oversight is specifically needed. Before committing to collaborate, users need transparency in the intentions of the autonomous agent. Collaborative systems have the affordance to direct people towards specific behaviors, such as pro-environmental awareness (e.g. [142]) and health and wellbeing (e.g. [60]). There is an ethical concern here for the users' ability to actually control their choices facing the persuasive power of agents [72]. Bratman [17] states that collaboration is a cooperation between intentional individuals each of whom sees and treats the other as such, and intending to coerce an individual in a certain direction bypasses his/her intentional agency. The aforementioned meshing of subplans is also needed in human-agent collaborations, where the subplans of the agent should be transparent and clear to users, and the users have assurance that the agent would work to empower their choices instead of prescribing actions.

Table 1: The 11 design considerations categorized by human-agent collaboration qualities, along with potential starting points for designers to address the design considerations

Collaboration qualities	Design considerations	Promising starting points for designers
Code of conduct	Are the agent's intentions and protocols visible to users?	Ensuring that the agent fulfills its task and moral responsibilities, and that the intentions and behavior protocols of the agent are transparent and clear to the users Ensuring that the agent is designed to be responsive to users' intentions, actions, and values Taking precautions for preventing negative behavior from users towards the agent
Task delegation	What task is the agent to perform?	Augmenting (and not replacing) the skills of humans, while levering on the agent's superpowers
Autonomy and control	What level of autonomy is appropriate for this agent?	Breaking the task into its components and assigning each to an agent, a human, or both, according to their skills and capabilities
	What kind of agent decisions do or do not require user awareness and approval?	Enabling the agent to read a social context, judge situations according to social norms, and identify emergencies
	How should an agent intervene?	Carefully designing the timing and form of proactive behaviors
	When and how to release or retain autonomy?	Creating a flexible and open system that users can adjust the agent's level of autonomy any time, flag exceptions in data, and opt-out from the decisions of the agent
Intelligibility	How to provide intelligibility into how an agent works and why it behaves in certain ways?	Integrating transparency into an agent's data sets and algorithms, as well as into its intentions, behaviors and uncertainties in its predictions
	How to explain the intentions and behaviors of agents?	Resorting to the psychological mechanisms of how humans explain things, while also considering potential communication means that are unique to agents Taking into consideration the specific explanation needs depending on the type of agent and context
Common ground	How to establish a common ground between the human and the agent?	Enabling the agent to read the users' intentions (e.g., through verbal commands, gaze, gestures) Embedding verbal and/or nonverbal communication means into the agent so that the users can understand the agent's status and intentions
Agent offering help	When and how an agent can offer help to humans?	Investigating the level of help proactivity requested by the users under specific contexts Enabling users to adjust the level of proactivity for help Resorting to human-to-human help offering strategies (e.g., Politeness Theory, discourse markers)
Agent requesting help	What are the most effective means for an agent to request help?	Investigating the factors that influence compliance from humans Designing agents that display the inherent uncertainty in the task/context

A similar transparency is also needed in the behavior protocols of the agent. Assuming that an individual has decided to collaborate with an agent by providing personal data, it is an important challenge to keep him/her in the loop and in control so as to determine how the service is provided and under which conditions [137]. In relation to voice activated personal agents that are shared in a household, for example, Luria et al. [96] asked intriguing questions such as "What should an agent do when a mother-in-law asks for her daughter-in-law's location? Should it share, prevaricate, stall, redirect the subject, or cynically refuse?". Answers to such questions should be clear to users before they commit to collaboration.

Agents are required to carry out their task and moral responsibilities, yet collaboration is a two-way street. Humans are also obliged to reciprocate by carrying out their responsibilities. In the last decade, several studies have observed bullying of robots [127], aggression towards autonomous vehicles [107], and harassment of conversational agents [19]. These behaviors obviously fall outside of good social conduct in collaborations and need to be avoided by humans. HRI field, for example, has been investigating what makes people more, or less, aggressive towards robots (e.g. [79]) and developing different strategies to stop aggression (e.g. [127, 141]). Here lies additional challenges for designers.

To sum, agents are required to be maximally sensitive not only to the end users' intentions and reasons for action, but also to societal norms and legal statuses [147]. In order to commit to collaborate with such agents, these sensitivities must be visible to users *(Design*

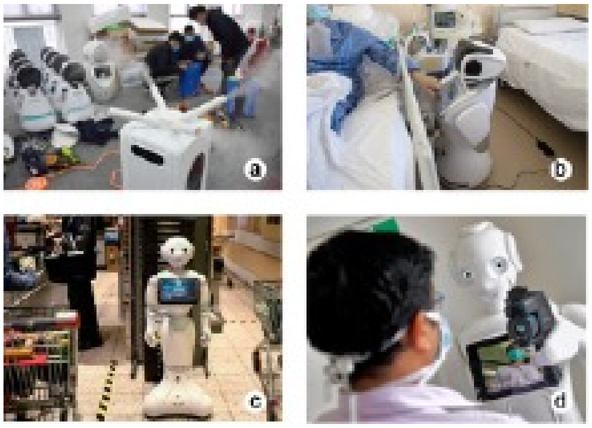


Figure 1: Some examples of robot use during the Covid-19 pandemic.

consideration 1: Are the agent's intentions and protocols visible to users?). In other words, people need to know the intentions (or subplans in Bratman's terms) of the agent and the expected behaviors depending on the embedded protocols in its system; and should be ensured that the system is designed to be responsive to their intentions, actions, and values. Designers, engineers, and programmers of these agents have an obligation to design the system in a way that this type of inquiry by users is not made impossible or unduly difficult [128]. They are also required to consider whether the agent would elicit any form of negative behavior from users, e.g., bullying, harassment, vandalism, and take necessary precautions in the design (or governance) of the agent.

3.1.2 Task delegation.

Design consideration 2: What task is the agent to perform? When individuals collaborate with agents, they willingly cede some of their decision-making power to technology. Affirming the principle of autonomy in the context of AI and agency means striking a balance between the decision-making power humans retain and that which it is delegated to agents [51]. A first step in this regard is to consider which tasks to delegate before committing to a joint task with an agent. A common narrative is that agents and robots are eminently suited for tasks that can be described with the three d's: dull, dirty, and dangerous [47]. At the height of the Covid-19 pandemic, for example, we have seen robots at work in different tasks (Figure 1). Some were utilized for spraying disinfectants in the hospitals and other indoor environments (Figure 1a), some checked in with the patients at the ICU (Figure 1b). Some robots were placed into public spaces such as supermarkets to recite the social distancing and hygiene rules (Figure 1c), and some others took people's temperature before entering buildings (Figure 1d).

Looking at it from the perspective of the three d's, the robots in 1a and 1b indeed freed humans from some dangerous tasks and allowed the care team to put their skills and efforts into other important tasks related to patient care. The common guideline to consider when designing agents is to always strive for this technology to support and enhance the skills of the human as opposed to substituting skills of the agents for skills of the human [81]. Humans and artificial agents have different capabilities. To take full advantage of human-agent collaboration, one must study how humans can most effectively augment agents and how agents can enhance what humans do best [158].

Some scholars in the HCI and HRI fields have already suggested to reflect more on what is unique to humans and unique to technology instead of blindly imitating humans in interactions. Welge and Hassenzahl [155], for instance, discussed various capabilities which are difficult for humans to maintain, while easy for robots to achieve provided by their particular "thingness". These are: being void of competitiveness, having endless patience, being unconditionally subordinated, having the ability to contain themselves, not taking things personally, and assuming responsibility. The authors named these social skills of robots as their "superpowers". The term superpowers indicate that each species, including the agent, serves its unique niche where its capabilities fit better than those of other species [43]. Similarly, Philips et al. [115] investigated human-animal teams as an analogue for human-robot teams. Animals have been actively domesticated to fulfill roles different than human based on their natural abilities, instincts, and physical capabilities. The authors maintain that human-robot collaborations are best facilitated through an approach in which the skills of robots complement the skills of humans, not unlike the ways in which humans and animals leverage skills to work interdependently. Going back to the robots in Figure 1c and 1d, we can argue that these robots actually released people from some dull and repetitive tasks as required by the three d's perspective. Yet, they do not exploit the full potential of superpowers related to being a (social) robot. The robot in 1c may as well be a sign or an announcement and the robot in 1d could be a thermal imaging camera. That is why their use comes across as gimmickier, compared to the examples in Figure 1a and 1b, which are better at exploiting their agentic potential.

Leveraging the unique capabilities of agents would also help to mitigate the disruptive effects of integrating agents into everyday life. For instance, in the HRI field, there is a vast amount of work investigating the introduction of robots in manufacturing settings, ranging from the technical to sociotechnical perspectives (e.g. [27, 114]). Welfare et al. [154] revealed the serious concern of workers regarding automation reducing the human interaction, movement and exercise, variation of tasks, possible autonomy, and physical work. By investigating which tasks to delegate to agents in specific contexts, employers can instead consider ways to augment employee wellbeing and productivity, while decreasing negative job attributes [98]. This augmentation does not necessarily have to be in rational, repetitive, and predictable tasks. Human-agent collaboration can happen in creative practices [94]. AI agents have been collaborating with musicians to compose music [101], supporting people in writing stories and slogans [33], helping actors to create introductions for improvised theatrical scenes [112], and supporting game designers to create levels in a game [164]. What is important in these cases is that the agents add to humans' existing abilities and keep the creativity and fun in the activity, instead of automating the task away entirely.

In summary, committing to a joint activity with an agent requires delegation of tasks. This delegation may occur more organically in human-to-human collaborations provided by the free will and consciousness of human beings, yet in the domain of artificial agents, deciding on who-does-what raises new considerations for designers *(Design consideration 2: What task is the agent to perform?)*. The tasks delegated to the agents should augment the skills of humans and not replace them; and understanding the user needs and the specific use context is critical, as in any design project, to identify these tasks. Second, designers also need to leverage the superpowers of agents. Instead of replicating attributes of humans, they need to explore the skills that agents can uniquely contribute for facilitating meaningful human-agent collaboration. This requires an exercise into overcoming the human frame of reference, take the best from "humanness" and "thingness" and combine it adequately into a meaningful interaction [155].

3.2 Mutual responsiveness

After delegating roles, the shared task execution commences in collaborations. As aforementioned, the collaborating parties are required to be mutually responsive during task execution, meaning that each party responds to what the other do and plan to do. In a similar vein, Johnson et al. [75] coined the term "coactive design" to characterize an approach in designing human-agent interactions that takes *interdependence* as the central organizing principle among humans and agents working together on a joint activity. Interdependence is meant to convey the reciprocal and mutually constraining nature of actions and effects that are conditioned by coordination [ibid].

Designing for interdependence has multiple facets. Imagine a war veteran who suffers from post-traumatic stress disorder (PTSD). He wears a smart jacket that helps him cope with stress and anxiety in everyday life. Borrowing this concept from [123], we can elaborate on it by imagining it as a partner in a collaboration. In this collaboration, the jacket tracks the heart rate and muscle tension of the wearer:

- When there is an increase, the jacket autonomously responds by either engaging the wearer with deep abdominal breathing to help him relax or fixating him when stress levels are too high to control (which may lead to irrational or aggressive behavior) [123]. The autonomy of the jacket can be overridden by the wearer. He can also activate the jacket any time he deems appropriate, e.g., as an anticipatory behavior before being exposed to human crowds.
- It is clear to him how the jacket works in terms of its inner sensing mechanism and behavior, so that he knows what and what not to expect when wearing it.
- The jacket is able to distinguish between stress behavior and other types of heart rate increasing activities, such as running for the bus; and the wearer is able to understand the momentary status and intentions of the jacket.

Taken altogether, these behaviors refer to three qualities of being mutually responsive, namely *autonomy and control*, *intelligibility*, and *common ground*, respectively, which will be discussed below in detail.

3.2.1 *Autonomy and control.*

Design consideration 3: What level of autonomy is appropriate for this agent? One of the important facets of mutual responsiveness in human-agent collaboration is about determining how much autonomy and control to give to agents. Beer et al. [10] defined the different levels of agent autonomy along the "sense, plan, act" primitives. An agent can vary in autonomy from low to high along these; and each of these primitives could be allocated to the human, the agent, or both. The PTSD jacket, for example, demonstrates high levels of autonomy in all of the primitives. When the jacket senses the increasing heart rate and tensing muscles, it plans on a strategy depending on the severity of the situation (e.g., giving biofeedback or fixating the person), and execute these behaviors. Packbot, on the other hand, is a partially tele-operated scout robot used in a military context for searching bomb-traps (Figure 2). Its capability of avoiding collisions with obstacles can be poor (i.e., low autonomy in sensing), and hence its human partner needs to intervene before the robot is seriously damaged (i.e., low autonomy in planning) [9]. Packbot then moves away from the potential obstacle autonomously (i.e., high autonomy in acting).

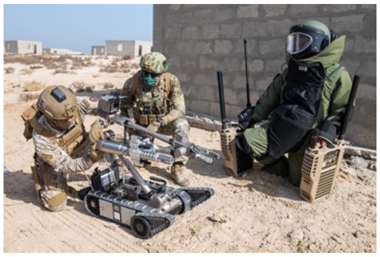


Figure 2: Packbot, a military search robot by Endeavor Robotics, has low autonomy in sensing and planning, but high autonomy in acting.

Most of the human-agent collaborations involve such a mixed-initiative control. Mixed-initiative systems requires a negotiation between the human and the agent, resulting in the system taking advantage of each of their skills, capabilities, and knowledge to best adapt to contingencies in the environment during task execution [90]. For instance, in relation to the control of Wilderness Search and Rescue robots, it was demonstrated that the human party focused resources better on likely search areas, but the robots tended to utilize large numbers of agents to systematically cover ground more effectively [66]. In order to support optimal human-agent collaboration, an agent's autonomy level and the level of control to be given to an individual should be carefully considered according to the task at hand and the specific skills and capabilities of the users *(Design consideration 3: What level of autonomy is appropriate for this agent?)*. Here, planning along the sense, plan, act primitives might be useful to break the tasks into components and consider which components to assign to agents and which to humans.

Design considerations 4 and 5: What kind of agent decisions do or do not require user awareness and approval? How should an agent intervene? Mixed-initiative control allows integrating the human into many complex control domains in the way the human often wants to be integrated—flexibly in charge and aware, but not required to issue every command [105]. Here lies additional design challenges as to when an agent should *take initiative* and when it is appropriate for an agent to *interrupt* the user. Luria et al. [96] demonstrated that the proactive interventions of a voice activated personal assistant must adapt to the social situations at the context of home. For example, the interventions of this agent were welcomed when it initiated a game with the boy as he was waiting for dinner, whereas felt uncomfortable when the parents made-up after a fight. Agents need to possess a capacity to decide on when to engage in a scene by understanding the social context, which seems not fully feasible in the near future. Another finding from the same study was that the participants expected the agent to recognize an emergency and report it [96]. This brings in more challenges related to being able to identify what counts as an emergency. One parent in the study, for example, expected the agent to be highly proactive if it learned that their teenager was drinking beer. This may count as an emergency in this particular household, but not in others. How an agent can judge situations according to (commonly established or localized) rules and norms is another difficult yet important question to explore.

Emergencies could be universally-agreed-upon cases for proactivity, but what about other interruptions? In general, interruptions are distracting, potentially leading to negative effects on task performance [99]. Prior research has specifically identified the appropriateness of the timing of an interruption as one of the most important factors dictating interruption consequences [121]. In this regard, Banerjee et al. [6] developed an algorithmic model for an agent to be able to assess the "interruptability degree" of a person. Regarding how an agent should interrupt, Saulnier et al. [129] demonstrated that people were able to interpret interruption urgency from robot behavior using only minimal nonverbal behavioral cues, such as motion speed and proximity to the person. Chiang et al [28] developed an online interactive learning framework for social robots to initiate an interruption in a socially acceptable manner by personalizing its behavior according to the attention of its partner and its conjecture about partner's awareness of itself.

Taken altogether, these insights indicate a necessity to carefully consider when an agent should intervene *(Design consideration 4: What kind of agent decisions do or do not require user awareness and approval?)*; and the timing and form of proactive behaviors *(Design consideration 5: How an agent should intervene?).* How an agent could read a social context, judge situations according to social rules and norms, and identify emergencies are challenges in this regard.

Design consideration 6: When and how to release or retain autonomy? Autonomy is generally conceived to be a fluid concept. Being integrated into the complexity and noise of the real-world usually render a static autonomy level insufficient to maintain high levels of performance. Here, flexible autonomy—also known in slightly different forms as dynamic task allocation, adjustable autonomy, or sliding autonomy—is key. Flexible autonomy refers to being able to dynamically change the autonomy level of a system and that a user can provide input into the system to improve its performance at any point in time [2]. In Bratman's collaboration framework, mutual responsiveness requires each individual to adjust his/her contributions in light of information about how the other party is faring. Flexible autonomy provides this adjustment in human-agent collaboration. Adjusting autonomy levels has known benefits for increasing the engagement of humans with the agents [77] and keeping a strong feeling of control over the agent [2]. Tariff Agent, for example, is a software agent that tracks energy consumption and energy tariffs, calculates the best tariff for the next day and (optionally) automatically switch to it [2]. It provides flexible autonomy by sending users updates about the status and decisions of the system and allowing them to switch easily between suggestion-only, semi-autonomous, and full autonomous options with an interface (Figure 3).

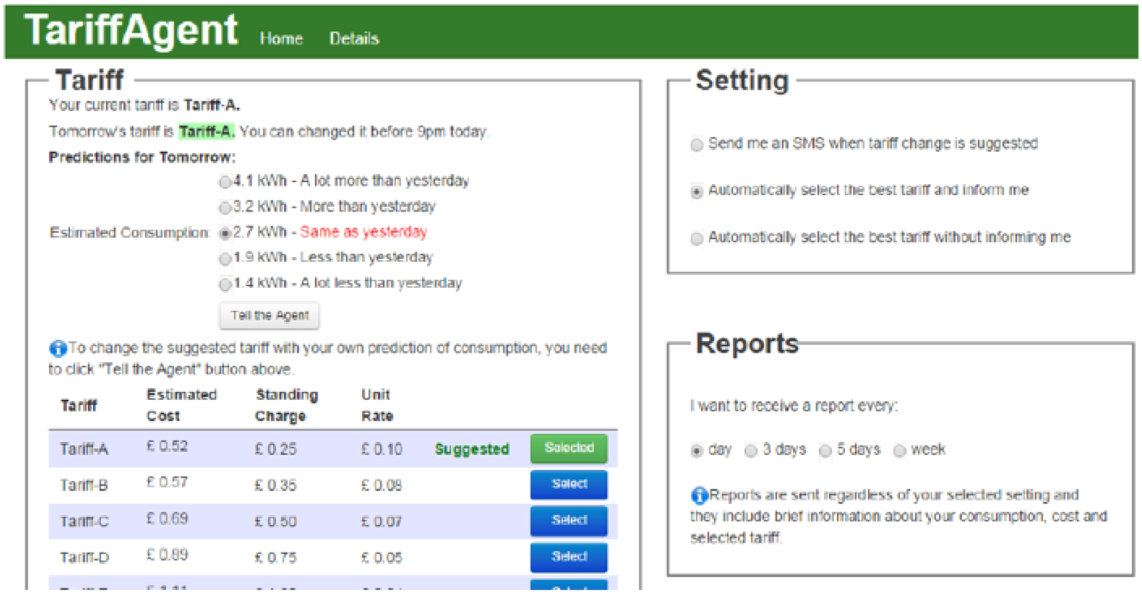
Being in control of the autonomy level also requires being able to override the autonomous decisions of an agent. Let us go back to the concept of "meshing of subplans" in collaborations. When the subplans of a user and an agent do not match after the collaboration has started, there must be options offered to the users in order to stop or re-plan the collaboration. The first option in this regard is to be able to flag exceptions. *Exception flagging* allows users to identify exceptional inputs that the system should ignore when building models and making predictions [163]. No matter how many sensors are included in an agent or how elaborate its algorithms become, there will be gaps in an agent's knowledge. Existing approaches to fill these gaps focus on giving feedback on the system's output or on eliciting more and higher quality input from the user [ibid]. Yang and Newman [163] demonstrated that none of these approaches seem well suited to an agent such as Nest, whose system-initiated changes become visible only after the change. The authors recommend an alternative approach which enables users to flag inputs that should not be learned. A key design challenge is, then, to create interfaces that can help the agent differentiate the data that represents regular, stable preferences from input that does not.

In addition to exception flagging, any delegation of control to an agent should remain overridable in principle [51]. Being able to opt-out from some or all of the decisions of an agent temporarily or permanently should be a key design consideration in collaborations. In relation to autonomous distributed energy systems, Pschetz et al. [118], for example, found that an agent that ruled out opportunities to negotiate faced a strong resistance from the participants. Explicit mechanisms for opting out have also been studied in relation to surveillance and wearable cameras [113, 117, 130] and privacy in teleoperation robots [22]. Creating mechanisms for users to flag exceptions or opt-out from the decisions of the agent should not be necessarily seen as removing or diminishing agency from the agent. As a matter of fact, the user input provides more data for the system to adjust its behavior, which is crucial for mutual responsiveness.

These insights bring us to the last design consideration in relation to autonomy and control, that is, leaving the system open to transfer of control by allowing users to adjust the system's level of autonomy any time they want, and being able to flag exceptions and opt-out from the decisions of the agent *(Design consideration 6: When and how to release or retain autonomy?).*

3.2.2 Intelligibility.

Design consideration 7: How to provide intelligibility into how an agent works and why it behaves in certain ways? The second requirement for being mutually responsive in human-agent collaboration is about helping users to understand how an agent interprets and acts upon the data it receives, i.e., *intelligibility.* The misalignment between users' mental models and the actual agent constraints can



Tariff

Your current tariff is Tariff-A.

Tomorrow's tariff is Tariff-A. You can changed it before 9pm today.

Predictions for Tomorrow:

Estimated Consumption:
○ 4.1 KWh - A lot more than yesterday
○ 3.2 KWh - More than yesterday
● 2.7 KWh - Same as yesterday
○ 1.9 KWh - Less than yesterday
○ 1.4 KWh - A lot less than yesterday

Tell the Agent

To change the suggested tariff with your own prediction of consumption, you need to click "Tell the Agent" button above.

Tariff	Estimated Cost	Standing Charge	Unit Rate		
Tariff-A	£ 0.52	£ 0.25	£ 0.10	Suggested	Selected
Tariff-B	£ 0.57	£ 0.35	£ 0.08		Select
Tariff-C	£ 0.69	£ 0.50	£ 0.07		Select
Tariff-D	£ 0.89	£ 0.75	£ 0.05		Select

Setting

○ Send me an SMS when tariff change is suggested

● Automatically select the best tariff and inform me

○ Automatically select the best tariff without informing me

Reports

I want to receive a report every:

● day ○ 3 days ○ 5 days ○ week

Reports are sent regardless of your selected setting and they include brief information about your consumption, cost and selected tariff.

Figure 3: Tariff Agent's homepage. The "setting" box allows for flexible autonomy (Courtesy of Alper Alan).

lead to over- or under-estimation of the agent's capabilities. Bringing back the two examples mentioned in the previous section, Alan et al. [2] had found that the participants perceived Tariff Agent's predictions to be smarter and more complex than they are in reality, even though the system's strategy was explicitly explained on the system's home page. Yang and Newman [163] demonstrated that the users of Nest overestimated its learning capabilities, which lead to some energy wastage. Overestimation was hazardous in these two cases; but it can lead to disastrous, even lethal, outcomes in critical systems and it should be avoided at all costs [57]. Maybe not as critical as overestimation, but under-estimation can also lead to negative consequences, such as a lack of communication and increased workload for humans. Under-estimation (and under-trust) can be detrimental to the effectiveness and efficiency of the collaborations since it may lead to disuse or micromanagement [41].

The primary focus of intelligibility research has been on providing explanations for how an agent works and why it behaves in certain ways. Central to these works is increasing the transparency of a system, such as its performance, process, intent, and purpose [3]. A major research stream on Explainable AI (XAI) focuses on all background information that is needed to make the decision-making processes of AI algorithms transparent and understandable, including the training data and model performance [110]. Abdul et al. [1], however, point to a delicate balance regarding explainability: XAI efforts typically focus on transforming complex "black box" models into simpler mathematically interpretable models. While these are significant contributions, these works tend to neglect whether they are usable and intelligible for humans. Making a system visible does not necessarily mean that it is understandable, which a challenging task when it comes to complex systems [137]. In this respect, transparency also brings in accessibility concerns and there is an opportunity for HCI research to bridge this gap [1].

There have been efforts in this direction by exploring transparency in conveying information about the uncertainty, dependence, and vulnerability of an AI system. Although many research go into creating robust models, currently uncertainty constitutes a fundamental attribute of any AI-driven system [14]. To mitigate for its impact, Intuitive Confidence Measure (ICM) model was designed to explain the likelihood of a correct single prediction based on its similarity to previous experiences, e.g. "I am *reasonably* certain that there is a victim at location A" [150]. Begoli et al. [11] recommended to implement a precise uncertainty estimation metric that incorporates all sources of decision uncertainty (e.g., performance on model capabilities, prior knowledge about the training data distribution, noises on input data) in software agents in order to emphasize the limitations of AI prediction. Jung et al. [76] operationalized this recommendation by designing an interface that displayed the uncertainty in an electric car's range estimation (Figure 4). This range estimation often leads to inaccurate predictions. While most of the prior work on car batteries focused on improving prediction accuracy by increasing the number of parameters that the system takes into account, Jung et al. [76] embraced the uncertainty and found that a range display that highlights prediction uncertainty led to improvements in driving experience, behavior and trust towards the car. Approaches such as this are in line with the advocacy in HCI to consider uncertainty and ambiguity as generative design material, rather than something to diminish (e.g. [14]).

Taken altogether, we can state that there are opportunities for HCI research to study how to align users' mental models of an agent's capabilities with actual constraints of the agent *(Design consideration 7: How to provide intelligibility into how an agent works and why it behaves in certain ways?)*. In this regard, designers might

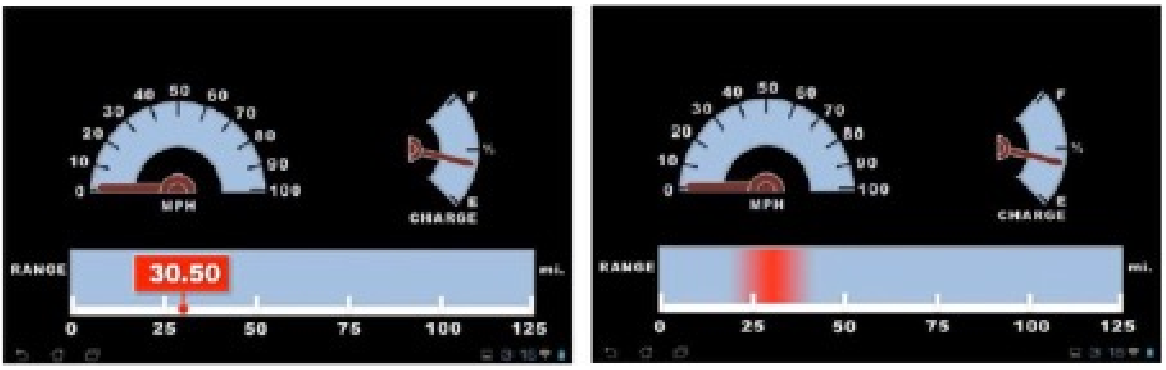


Figure 4: Remaining range was displayed as either highly precise values (left) or a diffuse color band (right) (Courtesy of Malte Jung).

experiment with integrating transparency into an agent's data sets, algorithms, and data models; as well as its intentions, behaviors, and uncertainty in its predictions.

Design consideration 8: How to explain the intentions and behaviors of agents? When it comes to "how" to explain, it is commonly assumed that a helpful explanation should align with how humans explain behavior [40, 104]. Harbers et al. [65], for example, demonstrated that users' explanations of an agent's behavior can be mapped to mental categories such as beliefs, desires/goals, intentions (BDI). The authors proposed to design agents according to the BDI principles in order to be explainable in human-compatible ways. Stange and Kopp [136] created a model that maps decision-making processes of a robot onto folk-psychological concepts found in human explanations: needs, intentions, actions. Their model includes causally structured explanations that connect two steps of the robot's reasoning process, e.g. "I did this because I intended that" or "I intended to do this because I needed that". The authors found that the explanations containing two elements with a marked causal relation were preferred over those containing only one, as they justified the behaviors better and increased their understandability. An interesting finding from the same study was that human-like needs-based explanations of robots (e.g. "I did this because I needed social contact or entertainment") led to a decrease in behavior desirability and lower ratings concerning how well they justified the behaviors, compared to intention- and action-based explanations which were related to the robot's rational agency.

This insight points to additional design challenges about investigating the levels of intentionality and desire that humans are willing to attribute to agents. Anthropomorphism is a delicate subject in human-agent interactions. While there are many examples of anthropomorphism providing a positive user experience (e.g. [46, 116]), there is also a recent stream of work that argue anthropomorphism is not the only solution to create good agents. Wortham and Rogers [161], for example, designed a robot that explains its behavior by a means that is both human and machine at the same time. The robot presents its generated plan by employing the social practice of muttering to oneself, but in a particular, machine-like way (i.e., "Attempting forward avoiding obstacle", "Doing sleep 10 seconds"). The design challenge here is to balance the product-ness, robot-ness, and humanlike-ness of an agent in understandable and creative ways [42] when it comes to explainability.

Yang and Newman [163] point out that many approaches to intelligibility and transparency assume that the users have a conscious interest in understanding the agent and are willing to invest time and energy in doing so. Their observations with Nest users suggested that the desire to understand the system arose only when something went wrong. However, it would be beneficial to the system's operation, and ultimately to the user, if they did see the value in understanding the agent's behavior. For this reason, Yang and Newman [163] recommend delivering intelligibility opportunistically, without requiring explicit interaction dedicated to the task. This *incidental intelligibility* is about designing interaction elements to increase the users' understanding of an agent's intelligent behavior in the tasks that the users consciously seek to accomplish, while avoiding asking users to learn how the system thinks as a discrete task [ibid]. Nest is a highly autonomous agent that allows small, occasional, and incidental interactions with users. On the other hand, when it comes to agents that interacts with users continuously or when the stakes are higher (e.g., AI-powered Clinical Decision Support Systems in hospitals or loan organization systems for mortgage), the willingness to understand the inner mechanics of an agent is likely to be higher. Explanation is a social process where humans tailor explanatory contents to different explainability needs [78]. The call for intelligibility and transparency can succeed only if we know which contexts require explaining and what forms of explanations are actually helpful.

These bring us to the second design consideration regarding intelligibility *(Design consideration 8: How to explain the intentions and behaviors of agents?)*. When giving explanations into agent behavior and intentions, designers might want to resort to psychological mechanisms that humans explain things in order to ensure the understandability of these explanations. However, while doing so, it is important to remember that agents might have unique means to communicate these mechanisms that combines humanlikeness and agent-ness, and users might have specific explanation needs depending on the context.

3.2.3 Common ground.

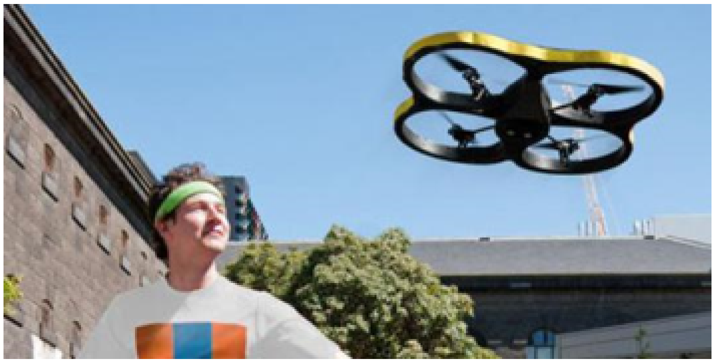


Figure 5: Joggobot, the running companion drone (Courtesy of Exertion Games Lab).

Design consideration 9: How to establish a common ground between the human and the agent? In collaborations, all the parties involved are in the context of the same joint activity. In addition to each agent having its own situation, there is also a need to be aware of the situation that the other parties are in. This collective understanding of context that is necessary for being mutually responsive brings in a new regulatory constraint that is referred to as *common ground* [75]. As defined by Clark and Brennan [36], common ground between two collaborators is the knowledge, beliefs, and suppositions they believe they share about the activity. Although the common ground framework was developed to understand communication and collaboration among people, recent work extended the framework into HCI and HRI (e.g. [25, 82, 138]). This research suggests that interfaces can be improved by framing the user's experience as a conversation in which a shared meaning between the user and the interface is developed [138]. This is, however, easier said than done in relation to agentic technologies. When two humans are co-present, they have the same perceptual access to the shared environment and generally do not have any problem understanding each other's references. Agents, on the other hand, have significantly mismatched perceptual and cognitive capabilities [24]. It is important for these parties to mediate differences and establish a common ground. This requires the agent to know the human's status, and likewise, the human to know the agent's status to the degree that they respond to each other's goals and actions [10].

To start with the former, let us consider the Joggobot [60]—a drone exercise companion (Figure 5). It flies in front of the runners to give them extra motivation or companionship during their run. To achieve these goals, Joggobot needs to know if the runner would want to slow down or speed up and which route they would like to take. In other words, it needs to correctly infer the intended goal of the runner in order to meaningfully assist them. This is known as the *intent inference problem* [73]. One approach to infer the user's intent could be to have the user explicitly communicate their intentions, such as via verbal commands. However, requiring explicit communication from the user could lead to increased cognitive load and ineffective collaboration [58]. Furthermore, giving verbal commands may be inappropriate in certain contexts for some people. A major stream of research in HRI investigate the use of nonverbal communication to infer intentions during collaborative tasks. Among many others, some examples include using users' *gaze patterns* to predict where an agent should look at [111] and to perform anticipatory actions [71]; utilizing *gestures* for object inference in interactions [156]; and using *motion patterns* for planning an agent's route [100].

When it comes to humans interpreting the agents' behavior, the concept of theory of mind (ToM) is an important starting point. ToM refers to an understanding that others have beliefs, desires, and intentions different from oneself, and that these beliefs, desires, and intentions influence their behavior (see [59] for a review of ToM). This mechanism allows people to take perspective of others and infer intentions and goals that underlie action [80]. Interacting with agents cannot help but give the users an impression that the agents act with beliefs, desires, and intentions [39]. This act of mind reading will be purely illusory, but the impression may nevertheless prove enduring if certain "tricks" are implemented in the agent [91]. Joggobot, for example, uses tricks such as mid-air twitch to signal the runner to start off or go faster [60].

One of the most investigated communication means is the *motion trajectories* of agents. For instance, it has been shown that motion alone is sufficient to convey intent and coordinate joint action [134] and that legible motion, planned to clearly express the robot's intent, leads to more seamless collaborations [44]. Sharma et al. [133] explored how the locomotion path of a flying quadrotor robot can communicate affect to people. Saerbeck and Bartneck [125] demonstrated that acceleration and curvature of robot motions had more significant impact in conveying different affect to humans than having facial features. Another means for communicating intent is the *gestures* of an agent. In this regard, *gaze* (e.g. [108]), *arm and body movements* (e.g., [85, 120]) were studied to express intent, states, and other relevant information. In the domain of non-anthropomorphic agents, Sirkin et al. [134] used a vertical lift movement in a robotic ottoman to suggest attention and readiness to move; Burneleit et al. [21] had the Impatient Toaster to shake nervously to remind the user to eat; and Yamaji et al. [162] designed a smart trash can that moved toward the trash with a twisting motion and vocal interaction to entice children to collect trash.

To sum, there is a need in human-agent interactions for the two parties to know each other's status and intentions so that the human and the agent could respond to each other *(Design consideration 9: How to establish a common ground between the human and the agent?)*. Designers are required to create this common ground by enabling the agent to read the users' intentions through explicit commands, gaze or gestures, and vice versa, i.e., embedding communication means into the agents in order to enable users to understand the agents' intentions. Exploring the richness of nonverbal communication is a promising direction to follow, considering both anthropomorphic and nonanthropomorphic agents.

3.3 Commitment to mutual support

The last characteristic of collaboration mentioned by Bratman is about the collaborating parties to provide support to each other when necessary. This includes behaviors such as providing guidance, suggestions or warning, and requesting assistance. Below, we will address these two behaviors of support under "agent offering help" and "agent requesting help".

3.3.1 Agent offering help.

Design consideration 10: When and how an agent can offer help to humans? Designing an agent that offers help requires to consider two aspects—when an agent offers help and how it offers help. To investigate the former, Baraglia et al. [8] proposed a developmentally motivated behavior in which the agent intervenes to help when it detects that the effects of a human's action were not as predicted. Sakita et al. [126] designed robot assistance behaviors triggered in different conditions, such as taking over when both hands of the user are occupied or providing verbal disambiguation when the user's hesitation is detected. Baraglia et al. [7] developed a joint task execution system for object manipulation that has three trigger mechanisms, namely human-initiated help (i.e., the user controls the timing of robot action), robot-initiated reactive help (i.e., the assistance is triggered when the robot detects that the user needs help), and robot-initiated proactive help (i.e., the robot helps whenever it can). Their results indicated that people collaborated best with a proactive robot when measured with team fluency metrics and preferred the proactive help over human-initiated or reactive help.

The foci of these studies are on task accomplishment where proactivity in helping is appreciated, yet when it comes to the helping behavior of all-round agents to be integrated in the complexity of the home environment, Luria et al. [96] presented a more nuanced picture. The authors demonstrated that the preferred help behavior of voice activated personal agents depended on the situation and context. Participants varied between wanting the agent to only be reactive to requests, be proactive by providing information, or be proactive by providing recommendations for a course of action. These preferences varied even within the preferences of each individual. While designing helping agents, Luria et al. [96] recommend thinking of proactivity in thresholds and identify when a threshold is crossed in interactions.

This brings us to "how" an agent could offer help. Research shows that negative human responses are common when an agent is perceived as issuing explicit commands or dictating task/interaction structure (e.g. [141]). As a strategy to mitigate the commanding tone implied in direct statements of advice and directions, Torrey et al. [145] proposed options for a robot's help-giving speech by drawing from Politeness Theory [20] and informal speech. These include using hedges (words like "I think" and "probably") and discourse markers (words like "I mean" and "like"). The authors demonstrated that when robots used hedges and discourse markers, they were perceived as more considerate and likeable, and less controlling. Nikolaidis et al. [109] compared two conditions for instruction giving, in which the robot explained to the human how it wanted to do a task and informed the human why it chose to act in a specific way. The authors showed that when the robot provided informative utterances, "I need to be able to see the door with my forward-facing camera", the adaptation rate of the humans to the robot increased significantly. This strategy is also in line with the aforementioned intelligibility of an agent.

Taken altogether, the proactivity of the help and the form of help have an immense impact on the users' experience with the agents *(Design consideration 10: When and how an agent can offer help to humans?)*. While proactive help might be appreciated when accomplishing some tasks, it may damage the whole experience under other situations. As in every design process, the needs of the users and the use context should be carefully examined in this regard beforehand. Furthermore, proactivity better be offered such as flexible autonomy, where users can adjust how much and when they need help in situ. When helping, agents could resort to human-to-human help offering strategies (i.e., using Politeness Theory or discourse markers) and/or give reasons of the instructions they give. It should be noted that these two examples involve verbal communication. What would be the equivalents of these polite behaviors in nonverbal agents is an interesting question to tackle for designers.

3.3.2 Agent requesting help.

Design consideration 11: What are the most effective means for an agent to request help? As agents move away from controlled environments into unstructured and dynamic environments, e.g., home, work, public space, they will almost always encounter situations that are beyond their capabilities. Engineers and computer scientists accept that any robot have many limitations and will continue to face various forms of limitations for the foreseeable future [151].

In most of these problematic situations, an agent may be more effective in achieving its tasks by proactively requesting assistance from people. In HRI, there are multiple studies which investigated the possibility of robots to augment their sensory or physical capabilities via a request for help. Michalowski et al. [103] designed a robot to play a game of "social tag" in which the task was to locate a team member who was wearing a pink hat. By asking "I am looking for the person in the pink hat. Can you help me?" to passers-by, the robot achieved its goal in 83% of the cases. Weiss et al. [153] revealed that an autonomous mobile robot could successfully navigate to a goal location in an outdoor scenario without any prior map or GPS, simply by asking for directions from bystanders. Rosenthal et al. [122] developed CoBot, which proactively requested help from humans in an office about localization, writing notes, and moving chairs. The authors found that people were willing to help the robot, even though it required some participants to interrupt their meetings and phone calls. Srinivasan and Takayama

[135] operationalized Politeness Theory [20] in scenarios that a robot requested help from people by using positive politeness, negative politeness, direct requests, and indirect requests. The authors demonstrated that using a positive politeness strategy (i.e., the help requester conveys liking toward the listener and makes them feel good about themselves, interests, or possessions; e.g. "You diagnose problems really well. If you teach me to diagnose the copier, it will be very helpful in the future") is more effective for robots that need help from people, especially when the robot is perceived as a peer.

In contrast to verbal help requests which were covered so far, another direction in the HRI field is to capitalize on more implicit nonverbal requests to elicit help from people. Yamaji [162], for example, demonstrated that movements and vocal cues can be used to effectively convey the purpose and required assistance of a minimalistic trash can robot. Gestures can be used for highlighting the uncertainty in an agent's action so that humans would take preemptive actions to ensure that the agent is able to accomplish the task. Van den Brule et al. [148] employed existing mechanisms of human signaling for uncertainty, e.g., scratching one's forehead, for alerting people to impending failures of a humanoid robot. The authors demonstrated that robots could solicit useful interventions from people to improve task performance by gesturally indicating a level of confidence or uncertainty about their actions. Kwon et al. [88] focused on the mechanisms for a robot to communicate when it is incapable of accomplishing a task. The authors designed expressive movements, which the robot executes a trajectory similar to the trajectory it would have executed had it been capable. Once the robot realizes that it is incapable of turning a valve, for example, it would find some motion (e.g., rotating its hand) that still conveys what the task is and sheds light on the cause of incapability. The authors found that these expressive movements improved not only the participants' understanding of what the robot is trying to do and why it cannot (i.e., intelligibility), but also their overall perception of the robot and willingness to collaborate with it.

The results of these studies indicate that people are willing to help agents when there is a need. Designers are required to envision ways to leverage this tendency of humans in the most effective ways *(Design consideration 11: What are the most effective means for an agent to request help?).* Investigating the factors that influence compliance [135] and designing agents that respond to the inherent uncertainty in the context [148] are promising starting points in this regard.

4 DISCUSSION

This article aims to identify the important issues to inform the future implementations of artificial agents that are supposed to work and dwell with humans. The three collaboration characteristics that were described in Bratman's Shared Cooperative Activity framework were used to organize the different lines of research prevalent in the HCI and HRI communities related to agentic technologies and identify the qualities that require attention for accomplishing effective and pleasant collaborations between humans and agents. These qualities were presented as 11 design considerations, which are about code of conduct, task delegation, autonomy and control, intelligibility, common ground, agent offering help and agent requesting help. With these design considerations, we hope to increase awareness about the different dimensions of human-agent collaborations, and to support other researchers in getting an overview about the current challenges in the HCI and HRI fields in relation to this topic, locating their work, and planning future research.

In the paper, different types of agents—from robots to smart products and software—were included. These are different product categories with varying characteristics and qualities. We consider that the 11 design considerations are relevant and sufficiently broad to apply to all these different agent types, yet the specifics of how these considerations could be met might differ from agent to agent. For instance, how the intentions and behaviors of a social robot could be communicated to users would probably be different than communicating the intentions of a smart tennis racket. There will even be additional legal constraints in this regard when it comes to a high-risk and highly-regulated agent such as an autonomous vehicle. Nevertheless, the necessity to take "intent communication" into account in the design of the collaboration is still present in these three different types of agents. That is why the design considerations were deliberately formalized as questions and kept broad: Designers could consider how to address them in their own ways according to the agent type and the design context. The hope is that the design considerations would stimulate and inform future research into the development of guidelines specific to different agent types.

The broad review of the relevant HCI and HRI work in this paper also allowed identifying some promising directions for addressing the design considerations. These were posed as starting points, rather than guidelines or heuristics, since being prescriptive about how to answer the design considerations was not the aim of this paper. The planned next steps in this research will be to use the design considerations and the proposed starting points in designing, prototyping, and evaluating collaborations. The ultimate goal is to arrive at validated "design patterns" [16] and best practices for human-agent collaboration, which would add to the recent stream of work on AI-related design guidelines (e.g., Amershi et al.'s [5] design guidelines for human-AI interaction and Hagendorff's [64] guidelines for AI ethics).

Using a framework that is intended for human-human collaborations to explain human-agent collaborations may seem to counteract posthumanism, which urges us to acknowledge that nonhumans, such as artificial agents, do not need to be similar to humans and should not be made similar. A post-humanistic perspective on AI enables AI to free itself from the burden to imitate the human and explore different kinds of being and intelligence [30]. A similar debate is ongoing in the HRI literature about the extent to which theories about human-human behavior could be transferred to human-robot behavior [48]. There is evidence that some characteristics of human-human interactions, such as perspective taking and common ground, are replicated in human-robot interaction (see for a review [86]); while some others, such as ascribing blame or expectations regarding task completion, do not (e.g. [68, 97]). Instead of aiming to replicate human-human interactions with human-robot interactions, Evers et al. [48] recommends comparing the latter with the former to fully explore and understand the boundary conditions for applying behavior theories to human-robot interaction and to develop more nuanced theories specific to human-robot interaction. This was also our aim in this paper. We used Bratman's framework on human-human collaboration not as a prescriptive approach on how human-agent collaborations should be, but more as a descriptive exercise into organizing the current work, distill insights from it, and identify research and design directions. These characteristics are fundamental to the efficacy and pleasantness of a collaboration, and even more so for attempts to design it between humans and agents.

Throughout the paper, we identified multiple differences between human-human and human-agent collaborations. Agents have physical and social superpowers which require exploiting in task delegation and establishing common ground. They fail a lot in the tasks they are assigned to, which require designing extra mitigation strategies. Although they have agency, the overseeing of their activities should always be a part of human responsibility. Furthermore, there are various behaviors that occur naturally in human-human collaborations, which either are impossible to emulate with the current technology (e.g. reading social cues; ascribing intentions, beliefs, and desires to others) or need to be explicitly programmed (e.g. behavior protocols regarding how an agent will cope with harassment, how it will deal with sensitive data, how it will intervene). All in all, there lies many practical and ethical challenges for designers.

The self-learning, proactive, and AI-infused agents are a different sort of artefact than their "non-connected" and "non-smart" counterparts. Hassenzahl et al. [67] argue that AI-driven technologies create a fundamental shift from an embodied relationship with technology (i.e., technology is perceived as an extensive of the self) to one of alterity (i.e., technology is perceived to be other). The authors entitle this class of intelligent agents as "otherware" and make a call-for-action to the HCI community to develop new models, software architectures, and design methods to design otherware properly. Correspondingly, Giaccardi and Redström [54] demonstrate that the issues that arise in relation to intelligence and data are no longer "centered" around the human. Human-centered approaches to design fail to support designers since what things are and how they come to present themselves as such are not primarily about functionality in the local sense anymore. They are about the interactions between humans and things, as well as between things and other things, without humans being aware of the exchanges taking place. For this reason, the authors explore what happens if one thinks of networked computational things not only as designed artefacts or technological enablers, but also as agents in a design space where they actually participate. In a similar vein, Wiltse [159] points out that intelligent things have lives of their own, to which humans are only partially privy as users. She suggests consulting to object-oriented ontology and alien phenomenology as analytic methods in order to account for the character and relations of things. In HRI, Alves-Oliveira et al. [4] critique the predominant discourse on robots which focus only on utilitarian values and robot ideals of autonomy and efficiency. The authors list various possible metaphors, such as neediness, domestication, destruction, and citizenship, as new frames-of-reference for how humans can utilize, relate to, and co-exist with robots. All these perspectives point to an urgent need to fundamentally reframe the conceptual space of how to design agents, if humans intend to shape this technology into desirable outcomes. We hope that the 11 design considerations would provide a starting point to unpack and envision the agents' role in collaborations in particular.

As aforementioned, in this paper we mainly focused on collaborations that involve a human and an agent, which are in line with Bratman's original "modest sociality" concept [17]. However, agents are part of a heterogenous system of connected actors, which involve other agents, humans, and the environment. Therefore, there are other possible configurations for collaboration. For example, *a human* can collaborate *with multiple agents*. This is a research area that is frequently addressed in the military context, such as search and rescue robot teams or combat systems with multiple UVs. Here, an operator collaborates with two to a dozen agents operating under swarm behavior. Transition of a human between different robots (i.e., when a robot in a heterogenous robot team lacks the capabilities to complete a task and summons another one who actually can) [140] and re-embodiment of an agent (i.e., an agent moves its social presence from one device to another, taking on the physical capabilities of each physical device) [119] are also interesting research areas in this configuration. Another collaboration configuration is between *multiple agents and multiple humans*, in which a team of operators manage a team of agents (e.g., Multi User-multi Robot System [MURS]). Lastly, *agent to agent* collaborations are also possible. Distributed AI systems (DAI), smart contracts, IoT blockchain applications, and robot-to-robot communication are some of the relevant research areas for this configuration.

Focusing solely on modest sociality in this paper may have limited giving a thorough overview of the complexity of each collaboration quality. For example, when one or more humans collaborate with multiple agents at the same time, the issue of autonomy and control becomes even more layered. There could be various classifications of control in these configurations: benevolent or competitive, team of hierarchical, static or shifting roles [32]. Another example is that the concept of intelligibility to become much more complex when it comes to agent-to-agent collaborations. Williams et al. [157] found that silent communications between robots in a team was perceived as being unnerving by co-located human teammates. In such collaboration configurations, designers are required to find new means for providing intelligibility. The main argument is that the design considerations that were presented in this paper about modest sociality would also apply to these different configurations of collaboration, yet their priority level and specificity would be different. Future work is needed to investigate how each collaboration quality plays out in different configurations, and eventually, create additional collaboration configuration-specific guidelines.

The idea of humans working with agents is complex and multifaceted, extending beyond technical boundaries to address ethical, societal, and philosophical issues. The big Vygotskian idea is that what makes human cognition different is not more individual brainpower, but rather the ability of humans to collaborate with others in collective activities [144]. Time has come for the agents to gain this skill. There are many challenges to address to move forward. We need to make sure that the agents perform useful and meaningful tasks, they are safe, transparent, intelligible, reliable, instructible, correctable, and negotiable. Ultimately, a key success factor for agents will be whether they are designed to work truly in concert with users [70].

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, Paper 582, 1–18. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3173574.3174156

[2] Alper T. Alan, Enrico Costanza, Sarvapali D. Ramchurn, Joel Fischer, Tom Rodden, and Nicholas R. Jennings. 2016. Tariff Agent: Interacting with a Future Smart Energy System at Home. *ACM Trans. Comput.-Hum. Interact.* 23, 4, Article 25 (September 2016), 28 pages. DOI:https://doi.org/10.1145/2943770

[3] Victoria Alonso and Paloma de la Puente. 2018. System transparency in shared autonomy: a mini review. Front Neurorobot 12, 83. DOI: https://doi.org/10.3389/fnbot.2018.00083

[4] Patrícia Alves-Oliveira, Maria Luce Lupetti, Michal Luria, Diana Löffler, Mafalda Gamboa, Lea Albaugh, Waki Kamino, Anastasia K. Ostrowski, David Puljiz, Pedro Reynolds-Cuéllar, Marcus Scheunemann, Michael Suguitan, and Dan Lockton. 2021. Collection of Metaphors for Human-Robot Interaction. Designing Interactive Systems Conference 2021. Association for Computing Machinery, New York, NY, USA, 1366–1379. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3461778.3462060

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, Paper 3, 1–13. DOI:10.1145/3290605.3300233

[6] Siddhartha Banerjee, Andrew Silva, and Sonia Chernova. 2018. Robot Classification of Human Interruptibility and a Study of Its Effects. *J. Hum.*-Robot Interact. 7, 2, Article 14 (October 2018), 35 pages. DOI:https://doi.org/10.1145/3277902

[7] Jimmy Baraglia, Maya Cakmak, Yukie Nagai, Rajesh Rao, and Minoru Asada. 2016. Initiative in robot assistance during collaborative task execution. In 2016 11th ACM/IEEE international conference on human-robot interaction (HRI), IEEE, 67-74. DOI: 10.1109/HRI.2016.7451735

[8] Jimmy Baraglia, Yukie Nagai, and Minoru Asada. 2014. Prediction error minimization for emergence of altruistic behavior. In 4th Intl. conf. on Development and Learning and on Epigenetic Robotics, IEEE, 281–286. DOI: 0.1109/DEVLRN.2014.6982994

[9] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-robot interaction: An introduction.* Chapter 10. Cambridge University Press.

[10] Jenay M. Beer, Arthur D. Fisk, and Wendy A. Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. *J. Hum.-Robot Interact.* 3, 2, 74-99. DOI: https://doi.org/10.5898/JHRI.3.2.Beer

[11] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. 2019. The need for uncertainty quantification in machine-assisted medical decision making. Nature-Machine Intelligence 1, 1 (jan 2019), 20–23. https://doi.org/10.1038/s42256-018-0004-1

[12] Rachel K. E. Bellamy, Sean Andrist, Timothy Bickmore, Elizabeth F. Churchill, and Thomas Erickson. 2017. Human-Agent Collaboration: Can an Agent be a Partner? In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (*CHI EA '17*). Association for Computing Machinery, New York, NY, USA, 1289–1294. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3027063.3051138

[13] Tony Belpaeme, James Kennedy, Paul Baxter, Paul Vogt, Emiel E. J. Krahmer, Stefan Kopp, Kirsten Bergmann, Paul Leseman, Aylin C. Kuntay, Tilbe Goksun, *et al.* 2015. L2tor-second language tutoring using social robots. In Proceedings of the ICSR 2015 WONDER Workshop. https://pub.uni-bielefeld.de/download/2900267/2900268.

[14] Jesse Josua Benjamin, Arne Berger, Nick Merrill, and James Pierce. 2021. Machine Learning Uncertainty as a Design Material: A Post-Phenomenological Inquiry. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, Article 171, 1–14. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3411764.3445481

[15] Olle Blomberg. 2015. Shared Agency: A Planning Theory of Acting Together By Michael Bratman. Analysis, 75, 2, 346. Oxford University Press. DOI:10.1093/analys/anu155

[16] Jan O. Borchers. 2008. A pattern approach to interaction design. In *Cognition, Communication and Interaction*, 114-131. Springer, London, 2008.

[17] Michael Bratman. 1992. Shared Cooperative Activity. *The Philosophical Review, 101,* 2, 327-341. doi:10.2307/2185537

[18] Michael Bratman. 1993. Shared intention. Ethics, 104, 97-113. DOI:https://doi-org.tudelft.idm.oclc.org/10.1086/293577

[19] Alfred Benedikt Brendel, Maike Greve, Stephan Diederich, Johannes Bührke, and Lutz M. Kolbe. 2020. You are an Idiot! How Conversational Agent Communication Patterns Influence Frustration and Harassment. 2020. AMCIS 2020 Proceedings. 13.

[20] Penelope Brown and Stephen C. Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, pp. 56-311. Cambridge University Press.

[21] Eva Burneleit, Fabian Hemmert, and Reto Wettach. 2009. Living interfaces: the impatient toaster. In *Proceedings of the 3rd International Conference on Tangible and Embedded Interaction* (*TEI '09*). Association for Computing Machinery, New York, NY, USA, 21–22. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/1517664.1517673

[22] Daniel J. Butler, Justin Huang, Franziska Roesner, and Maya Cakmak. 2015. The Privacy-Utility Tradeoff for Remotely Teleoperated Robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (*HRI '15*). Association for Computing Machinery, New York, NY, USA, 27–34. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/2696454.2696484

[23] Jaime R. Carbonell. 1970. Mixed-initiative man-computer instructional dialogues. Final report. Tech. Rep. No. BBN-1971, Job No. 11399. Cambridge, MA: Bolt Beranek & Newman, Inc.

[24] Elizabeth Cha, Anca D. Dragan, and Siddhartha S. Srinivasa. 2015. Perceived robot capability. In Proceedings of the Twenty-Fourth IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). 541–548. DOI: 10.1109/ROMAN.2015.7333656

[25] Joyce Y. Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue." In 2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 33-40.

[26] Mai Lee Chang, Taylor Kessler Faulkner, Thomas Benjamin Wei, Elaine Schaertl Short, Gokul Anandaraman, and Andrea Lockerd Thomaz. 2020. TASC: Teammate Algorithm for Shared Cooperation. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE. DOI: 10.1109/IROS45743.2020.9340983

[27] EunJeong Cheon and Norman Makoto Su. 2016. Integrating roboticist values into a Value Sensitive Design framework for humanoid robots. 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 375-382, doi: 10.1109/HRI.2016.7451775.

[28] Yi-Shiu Chiang, Ting-Sheng Chu, Chung Dial Lim, Tung-Yen Wu, Shih-Huan Tseng, and Li-Chen Fu. 2014. Personalizing robot behavior for interruption in social human-robot interaction. In 2014 IEEE International Workshop on Advanced Robotics and its Social Impacts. IEEE, 44–49. https://doi.org/10.1109/ARSO.2014.7020978

[29] Nazli Cila, Iskander Smit, Elisa Giaccardi, and Ben Kröse. 2017. Products as Agents: Metaphors for Designing the Products of the IoT Age. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 448–459. DOI:https://doi.org/10.1145/3025453.3025797

[30] Mark Coeckelbergh. 2020. AI Ethics. MIT Press.

[31] Philip R. Cohen and Hector J. Levesque. 1991. Teamwork. Nous 25, 4, 487-512.

[32] Francesco Corea. 2019. Distributed Artificial Intelligence. Retrieved from: https://francesco-ai.medium.com/distributed-artificial-intelligence-3e3491e0771c

[33] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In 23rd International Conference on Intelligent User Interfaces (IUI' 18). Association for Computing Machinery, New York, NY, USA, 329–340. DOI:http://dx.doi.org/10.1145/3172944.3172983

[34] Herbert H. Clark. 1996. Using language. New York, NY: Cambridge University Press.

[35] Herbert H. Clark. 2005. Coordinating with each other in a material world. Discourse Studies, 7, 507–525.

[36] Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), Perspectives on Socially Shared Cognition (127-149). Washington, D.C.: American Psychological Association.

[37] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: machines must learn to find common ground. Nature, 593, 33-36. DOI: 10.1038/d41586-021-01170-0

[38] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. 2020. Impact of Agent Reliability and Predictability on Trust in Real Time Human-Agent Collaboration. In *Proceedings of the 8th International Conference on Human-Agent Interaction* (*HAI '20*). Association for Computing Machinery, New York, NY, USA, 131–139. DOI:https://doi.org/10.1145/3406499.3415063

[39] Maartje M. A. de Graaf and Bertram F. Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). AAAI 2017 Fall Symposium on "AI-HRI", 19-26.

[40] Maartje M. A. de Graaf and Bertram F. Malle. 2019. People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 239-248. doi: 10.1109/HRI.2019.8673308.

[41] Ewart J. De Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. Int. j. of soc. robotics 12, 2, 459-478. DOI:10.1007/s12369-019-00596-x

[42] Carl F. Disalvo, Francine Gemperle, Jodi Forlizzi, and Sara Kiesler. 2002. All robots are not created equal: The design and perception of humanoid robot heads. In Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (DIS'02). 321–326.

[43] Judith Dörrenbächer, Diana Löffler, and Marc Hassenzahl. 2020. Becoming a Robot - Overcoming Anthropomorphism with Techno-Mimesis. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–12. DOI:https://doi.org/10.1145/3313831.3376507

[44] Anca D. Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S. Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In Proceedings of the 10th Annual ACM/IEEE International Conference on Human–Robot Interaction. ACM, 51–58.

[45] Alpana Dubey, Kumar Abhinav, Sakshi Jain, Veenu Arora, and Asha Puttaveerana. 2020. HACO: A Framework for Developing Human-AI Teaming. In *Proceedings of the 13th Innovations in Software Engineering Conference on Formerly known as India Software Engineering Conference* (*ISEC 2020*). Association for Computing Machinery, New York, NY, USA, Article 10, 1–9. DOI:10.1145/3385032.3385044

[46] Brian Duffy. 2003. Anthropomorphism and the social robot. 2003. Robotics and Autonomous Systems, 42, 3, 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

[47] Joseph F. Engelberger. 1980. Robotics in practice: management and applications of industrial robots. Springer Science & Business Media.

[48] Vanessa Evers, Heidy Maldonado, Talia Brodecki, and Pamela Hinds. 2008. Relational vs. group self-construal: Untangling the role of national culture in HRI. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 255-262. doi: 10.1145/1349822.1349856.

[49] Umer Farooq and Jonathan Grudin. 2016. Human-computer integration. *interactions* 23, 6 (November-December 2016), 26–32. DOI:https://doi.org/10.1145/3001896

[50] Fanny Ficuciello, Guglielmo Tamburrini, Alberto Arezzo, Luigi Villani, and Bruno Siciliano. 2019. Autonomy in surgical robots and its meaningful human control. J. of Behavioral Robotics 10, 1, 30-43. DOI: 10.1515/pjbr-2019-0002

[51] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge *et al.* 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28, 4, 689-707. DOI:10.1007/s11023-018-9482-5

[52] Terrence W. Fong. 2001. Collaborative control: A robot-centric model for vehicle teleoperation. Pittsburgh, PA: Robotics Institute, Carnegie Mellon University.

[53] Eduard Fosch-Villaronga, Pranav Khanna, Hadassah Drukarch, and Bart H. M. Custers. 2021. A human in the loop in surgery automation. Nature Machine Intelligence 3, 5, 368-369. DOI:10.1038/s42256-021-00349-4

[54] Elisa Giaccardi and Johan Redström. 2020. Technology and more-than-human design. Design Issues 36, 4, 33-44.

[55] Margaret Gilbert. 1987. Modelling collective belief. Synthese, 73, 1, 185-204.

[56] Margaret Gilbert. 2003. The structure of the social atom: Joint commitment as the foundation of human social behavior. In Socializing metaphysics, ed. Frederick Schmitt. Lanham, MD: Rowman & Littlefield. 39-64.

[57] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J. of the American Medical Informatics Association, 19, 1, 121-127. DOI:10.1136/amiajnl-2011-000089

[58] Michael A. Goodrich and Dan R. Olsen. 2003. Seven principles of efficient human robot interaction. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. DOI: 10.1109/ICSMC.2003.1244504

[59] Alvin I. Goldman. 2012. Theory of mind. In E. Margolis, & S. P. Stich (Eds.). *The Oxford Handbook of Philosophy of Cognitive Science*, 402–424. London: Oxford University Press.

[60] Eberhard Graether and Florian Mueller. 2012. Joggobot: a flying robot as jogging companion. *CHI '12 Extended Abstracts on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1063–1066. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/2212776.2212386

[61] Barbara J. Grosz. 1996. Collaborative systems (AAAI-94 presidential address). *AI magazine,* 17, 2, 67.

[62] Barbara J. Grosz and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence,* 86, 2, 269-357.

[63] Barbara J. Grosz and Candace L. Sidner. 1990. Plans for discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), Intentions in communication (Chapter 20). Cambridge, MA: The MIT Press.

[64] Thilo Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds & Machines, 30, 99–120. DOI: 10.1007/s11023-020-09517-8

[65] Maaike Harbers, Karel van den Bosch, and John-Jules Ch. Meyer. 2009. A study into preferred explanations of virtual agent behavior. In Proceedings of the 9th International Conference on Intelligent Virtual Agents. Amsterdam, The Netherlands, 132-145. https://doi.org/10.1007/978-3-642-04380-2_17

[66] Benjamin Hardin and Michael A. Goodrich. 2009. On using mixed-initiative control: a perspective for managing large-scale robotic teams. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction* (*HRI '09*). Association for Computing Machinery, New York, NY, USA, 165–172. DOI:https://doi.org/10.1145/1514095.1514126

[67] Marc Hassenzahl, Jan Borchers, Susanne Boll, Astrid Rosenthal-von der Pütten, and Volker Wulf. 2020. Otherware: how to best interact with autonomous systems. interactions 28, 1 (January - February 2021), 54–57. DOI:https://doi.org/10.1145/3436942

[68] Pamela J. Hinds, Teresa L. Roberts, and Hank Jones. 2004. Whose job is it anyway? A study of human-robot interaction in a collaborative task. Human–Computer Interaction, 19, 1-2, 151-181. DOI: 10.1080/07370024.2004.9667343

[69] Guy Hoffman and Cynthia Breazeal. 2004. Collaboration in human-robot teams. In AIAA 1st Intelligent Systems Technical Conference. 6434.

[70] Lars Erik Holmquist. 2017. Intelligence on tap: artificial intelligence as a new design material. *interactions* 24, 4 (July-August 2017), 28–33. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3085571

[71] Chien-Ming Huang and Bilge Mutlu. 2016. Anticipatory robot control for efficient human-robot collaboration. In Proceedings of 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16). DOI:10.1109/HRI.2016.7451737

[72] Giulio Jacucci, Anna Spagnolli, Jonathan Freeman, and Luciano Gamberini. 2015. Symbiotic interaction: a critical definition and comparison to other human-computer paradigms. In International workshop on symbiotic interaction, pp. 3-20. Springer, Cham.

[73] Siddarth Jain and Brenna Argall. 2019. Probabilistic Human Intent Recognition for Shared Autonomy in Assistive Robotics. J. Hum.-Robot Interact. 9, 1, Article 2 (December 2019), 23 pages. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3359614

[74] Nicholas R. Jennings, Luc Moreau, David Nicholson, Sarvapali Ramchurn, Stephen Roberts, Tom Rodden, and Alex Rogers. 2014. Human-agent collectives. Communications of the ACM 57, 12, 80-88. DOI:https://doi.org/10.1145/2629559

[75] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna van Riemsdijk, and Maarten Sierhuis. 2014. Coactive design: designing support for interdependence in joint activity. J. Hum.-Robot Interact. 3, 1 (February 2014), 43–69. DOI:https://doi.org/10.5898/JHRI.3.1.Johnson

[76] Malte F. Jung, David Sirkin, Turgut M. Gür, and Martin Steinert. 2015. Displayed Uncertainty Improves Driving Experience and Behavior: The Case of Range Anxiety in an Electric Car. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 2201–2210. DOI:10.1145/2702123.2702479

[77] David B. Kaber and Mica R. Endsley. 2004. The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. Theoretical Issues in Ergonomics Science, 5, 2, 113-153. DOI:10.1080/1463922021000054335

[78] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). Lisbon, Portugal, 676ś682. https://doi.org/10.1109/ROMAN.2017.8172376

[79] Merel Keijsers, Hussain Kazmi, Friederike Eyssel, and Christoph Bartneck. 2021. Teaching Robots a Lesson: Determinants of Robot Punishment. *Int J of Soc Robotics* 13**,** 41–54. DOI:10.1007/s12369-019-00608-w

[80] Richard Kelley, Alireza Tavakkoli, Christopher King, Monica Nicolescu, Mircea Nicolescu, and George Bebis. 2008. Understanding human intentions via hidden markov models in autonomous mobile robots. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction* (*HRI '08*). Association for Computing Machinery, New York, NY, USA, 367–374. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/1349822.1349870

[81] Paul T. Kidd, Mansour Rahimi, and Waldemar Karwowski. 1992. Design of human-centered robotic systems. In *human robot interaction*, pp. 225-241. London, UK: Taylor and Francis.

[82] Sara Kiesler. 2005. Fostering common ground in human-robot interaction. In ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, IEEE, 729-734. DOI:10.1109/ROMAN.2005.1513866

[83] Peter Y. Kim. 2020. Where We're Going, We Don't Need Drivers: Autonomous Vehicles and AI-Chaperone Liability. Cath. UL Rev. 69, 341.

[84] Günther Knoblich, Stephen Butterfill, and Natalie Sebanz. 2011. Psychological research on joint action: theory and data. Psychology of learning and motivation, 54, 59-101.

[85] Hideki, Kozima, Marek P. Michalowski, and Cocoro Nakagawa. 2009. Keepon. Int. J. of Soc. Robotics, 1, 1, 3-18.

[86] Nicole C. Krämer, Astrid von der Pütten, and Sabrina Eimler. 2012. Human-agent and human-robot interaction theory: similarities to and differences from human-human interaction. In Human-computer interaction: The agency perspective, pp. 215-240. Springer, Berlin, Heidelberg.

[87] Christopher Kutz. 2000. Acting together. Philosophical and Phenomenological Research, 1-31.

[88] Minae Kwon, Sandy H. Huang, and Anca D. Dragan. 2018. Expressing Robot Incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (*HRI '18*). Association for Computing Machinery, New

York, NY, USA, 87–95. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3171221.3171276

[89] Maria Kyrarini, Fotios Lygerakis, Akilesh Rajavenkatanarayanan, Christos Sevastopoulos, Harish R. Nambiappan, Kodur K. Chaitanya, Ashwin R. Babu, Joanne Mathew, and Fillia Makedon. 2021. A Survey of Robots in Healthcare. *Technologies,* 9, 1, 8. https://doi.org/10.3390/technologies9010008

[90] David Landén, Fredrik Heintz, and Patrick Doherty. 2010. Complex task allocation in mixed-initiative delegation: A UAV case study. In International conference on principles and practice of multi-agent systems, 288-303. Springer, Berlin, Heidelberg.

[91] Florent Levillain and Elisabetta Zibetti. 2017. Behavioral objects: the rise of the evocative machines. J. Hum.-Robot Interact. 6, 1 (May 2017), 4–24. DOI:10.5898/JHRI.6.1.Levillain

[92] Joseph C.R. Licklider. 1960. Man-computer symbiosis. IRE transactions on human factors in electronics, 1, 4-11.

[93] Amanda Loudin. 2019. Novak Djokovic Used A.I. to Train for Wimbledon. Retrieved Sep 8, 2021 from https://onezero.medium.com/novak-djokovic-used-a-i-to-train-for-wimbledon-e22b62f7504a

[94] Todd Lubart. 2005. How can computers be partners in the creative process: classification and commentary on the special issue. Int. J. of Human-Computer Studies, 63, 4-5, 365–369.

[95] Kirk Ludwig. 2020. What Is Minimally Cooperative Behavior?. In Minimal Cooperation and Shared Agency, 9-39. Springer, Cham.

[96] Michal Luria, Rebecca Zheng, Bennett Huffman, Shuangni Huang, John Zimmerman, and Jodi Forlizzi. 2020. Social Boundaries for Personal Agents in the Interpersonal Space of the Home. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–12. DOI:https://doi.org/10.1145/3313831.3376311

[97] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 117-124.

[98] Maja J. Matarić. 2017. Socially assistive robotics: Human augmentation versus automation. Science Robotics, 2, 4.

[99] Daniel McFarlane and Kara Latorella. 2002. The scope and importance of human interruption in human-computer interaction design. Human-Computer Interaction 17, 1 (Mar. 2002), 1–61. https://doi.org/10.1207/S15327051HCI1701_1

[100] Jim Mainprice and Dmitry Berenson. 2013. Human-robot collaborative manipulation planning using early prediction of human motion. In 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 299-306. DOI:0.1109/IROS.2013.6696368

[101] Jon McCormack, Toby Gifford, Patrick Hutchings, Maria Teresa Llano Rodriguez, Matthew Yee-King, and Mark d'Inverno. 2019. In a Silent Way: Communication Between AI and Improvising Musicians Beyond Sound. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Paper 38, 1–11. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3290605.3300268

[102] Giulio Mecacci and Filippo Santoni de Sio. 2020. Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. Ethics and Information Technology, 22, 2, 103-115. DOI:10.1007/s10676-019-09519-w

[103] Marek P. Michalowski, Selma Šabanović, Carl DiSalvo, Didac Busquets, Laura M. Hiatt, Nik A. Melchior, and Reid Simmons. 2007. Socially distributed perception: Grace plays social tag at aaai 2005. Autonomous Robots, 22, 4, 385-397. DOI: 10.1007/s10514-006-9015-6

[104] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

[105] Christopher A. Miller and Raja Parasuraman. 2007. Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. Human factors, 49, 1, 57-75. DOI:10.1518/001872007779598037

[106] Tal Montal and Zvi Reich. 2017. I, robot. You, journalist. Who is the author? Authorship, bylines and full disclosure in automated journalism. Digital journalism, 5, 7, 829-849. DOI:10.1080/21670811.2016.1209083

[107] Dylan Moore, Rebecca Currano, Michael Shanks, and David Sirkin. 2020. Defense Against the Dark Cars: Design Principles for Griefing of Autonomous Vehicles. Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery, New York, NY, USA, 201–209. DOI:10.1145/3319502.3374796

[108] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In Proceedings of the 4th ACM/IEEE international conference on Human robot interaction (HRI '09). Association for Computing Machinery, New York, NY, USA, 61–68. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/1514095.1514109

[109] Stefanos Nikolaidis, Minae Kwon, Jodi Forlizzi, and Siddhartha Srinivasa. 2018. Planning with Verbal Communication for Human-Robot Collaboration. J. Hum.-Robot Interact. 7, 3, Article 22 (December 2018), 21 pages. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3203305

[110] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. UserModeling and User-Adapted Interaction 27, 3-5 (2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

[111] Dimitri Ognibene and Yiannis Demiris. 2013. Towards active event recognition. In Twenty-Third International Joint Conference on Artificial Intelligence.

[112] Brian O'Neill, Andreya Piplica, Daniel Fuller, and Brian Magerko. 2011. A knowledge-based framework for the collaborative improvisation of scene introductions. In International Conference on Interactive Digital Storytelling, 85-96. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-25289-1_10

[113] Shwetak N. Patel, Jay W. Summet, and Khai N. Truong. 2009. Blindspot: Creating capture-resistant spaces. In *Protecting Privacy in Video Surveillance*, pp. 185-201. Springer, London. DOI: 10.1007/978-1-84882-301-3_11

[114] Q. C. Pham, R. Madhavan, L. Righetti, W. Smart, and R. Chatila. 2018. The impact of robotics and automation on working conditions and employment. IEEE Robotics & Automation Magazine 25, 2, 126-128. DOI:10.1109/mra.2018.2822058

[115] Elizabeth Phillips, Kristin E. Schaefer, Deborah R. Billings, Florian Jentsch, and Peter A. Hancock. 2016. Human-animal teams as an analog for future human-robot teams: influencing design and fostering trust. J. Hum.-Robot Interact. 5, 1 (March 2016), 100–125. DOI:https://doi.org/10.5898/JHRI.5.1.Phillips

[116] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F. Malle. 2018. What is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18). ACM, New York, NY, USA, 105–113. https://doi.org/10.1145/3171221.3171268

[117] James Pierce. 2019. Smart Home Security Cameras and Shifting Lines of Creepiness: A Design-Led Inquiry. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Paper 45, 1–14. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3290605.3300275

[118] Larissa Pschetz, Kruakae Pothong, and Chris Speed. 2019. Autonomous Distributed Energy Systems: Problematising the Invisible through Design, Drama and Deliberation. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Paper 387, 1–14. DOI:https://doi.org/10.1145/3290605.3300617

[119] Samantha Reig, Michal Luria, Janet Z. Wang, Danielle Oltman, Elizabeth Jeanne Carter, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2020. Not Some Random Agent: Multi-person Interaction with a Personalizing Service Robot. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction.* Association for Computing Machinery, New York, NY, USA, 289–297. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3319502.3374795

[120] Laurel D. Riek, Tal-Chen Rabinowitch, Paul Bremner, Anthony G. Pipe, Mike Fraser, and Peter Robinson. 2010. Cooperative gestures: Effective signaling for humanoid robots. In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 61-68. DOI:10.1109/HRI.2010.5453266

[121] A. Joy Rivera. 2014. A socio-technical systems approach to studying interruptions: Understanding the interrupter's perspective. Applied Ergonomics 45, 3 (May 2014), 747–756. DOI:https://doi.org/10.1016/j.apergo.2013.08.009

[122] Stephanie Rosenthal, Manuela Veloso, and Anind K. Dey. 2012. Is someone in this office available to help me?. J. of Intelligent & Robotic Systems, 66, 1, 205-221. DOI:10.1007/s10846-011-9610-4

[123] Marco Rozendaal. 2016. Objects with intent: a new paradigm for interaction design. *interactions* 23, 3 (May + June 2016), 62–65. DOI:https://doi.org/10.1145/2911330

[124] Marco C. Rozendaal, Boudewijn Boon, and Victor Kaptelinin. 2019. Objects with Intent: Designing Everyday Things as Collaborative Partners. ACM Trans. Comput.-Hum. Interact. 26, 4, Article 26 (July 2019), 33 pages. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3325277

[125] Martin Saerbeck and Christoph Bartneck. 2010. Perception of affect elicited by robot motion. In *2010 5th ACM/IEEE International* Conference on Human-Robot Interaction (HRI), IEEE, 53-60. DOI:10.1109/HRI.2010.5453269

[126] Kenji Sakita, Koichi Ogawara, Shinji Murakami, Kentaro Kawamura, and Katsushi Ikeuchi. 2004. Flexible cooperation between human and robot by interpreting human intention from gaze information. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566), vol. 1, IEEE, 846-851. DOI:10.1109/IROS.2004.1389458.

[127] Pericle Salvini, Gaetano Ciaravella, Wonpil Yu, Gabriele Ferri, Alessandro Manzi, Barbara Mazzolai, Cecilia Laschi, Sang-Rok Oh, and Paolo Dario. 2010. How safe are service robots in urban environments? Bullying a robot. In *19th international symposium in robot and human interactive communication*, IEEE, 1-7. DOI:10.1109/ROMAN.2010.5654677

[128] Santoni de Sio, Filippo, and Jeroen Van den Hoven. "Meaningful human control over autonomous systems: A philosophical account." Frontiers in Robotics and AI 5 (2018): 15. DOI:10.3389/frobt.2018.00015

[129] Paul Saulnier, Ehud Sharlin, and Saul Greenberg. 2011. Exploring minimal nonverbal interruption in HRI. In 2011 RO-MAN. IEEE, 79–86. https://doi.org/10.1109/ROMAN.2011.6005257

[130] Jeremy Schiff, Marci Meingast, Deirdre K. Mulligan, Shankar Sastry, and Ken Goldberg. 2009. Respectful cameras: Detecting visual markers in real-time to address privacy concerns. In Protecting Privacy in Video Surveillance, 65-89. Springer, London. DOI:10.1007/978-1-84882-301-3_5

[131] Natalie Sebanz, Harold Bekkering, and Günther Knoblich. 2006. Joint action: bodies and minds moving together. Trends in cognitive sciences, 10, 2, 70-76. DOI: 10.1016/j.tics.2005.12.009

[132] Bernhard Sendhoff and Heiko Wersing. 2020. Cooperative Intelligence-A Humane Perspective. In 2020 IEEE International Conference on Human-Machine Systems (ICHMS), IEEE, 1-6. doi: 10.1109/ICHMS49158.2020.9209387.

[133] Megha Sharma, Dale Hildebrandt, Gem Newman, James E. Young, and Rasit Eskicioglu. 2013. Communicating affect via flight path exploring use of the laban effort system for designing affective locomotion paths. In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 293-300. doi: 10.1109/HRI.2013.6483602.

[134] David Sirkin, Brian Mok, Stephen Yang, and Wendy Ju. 2015. Mechanical Ottoman: Up Close and Personal. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts (HRI'15 Extended Abstracts). Association for Computing Machinery, New York, NY, USA, 297. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/2701973.2714391

[135] Vasant Srinivasan and Leila Takayama. 2016. Help Me Please: Robot Politeness Strategies for Soliciting Help From Humans. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 4945–4955. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/2858036.2858217

[136] Sonja Stange and Stefan Kopp. 2020. Effects of a Social Robot's Self-Explanations on How Humans Understand and Evaluate Its Behavior. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, 619–627. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3319502.3374802

[137] Constantine Stephanidis, Gavriel Salvendy, Margherita Antona, Jessie YC Chen, Jianming Dong, Vincent G. Duffy, Xiaowen Fang *et al.* 2019. Seven HCI grand challenges. Int. J. of Human–Computer Interaction, 35, 14, 1229-1269. DOI:10.1080/10447318.2019.1619259

[138] Kristen Stubbs, David Wettergreen, and Illah Nourbakhsh. 2008. Using a robot proxy to create common ground in exploration tasks. In 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 375-382. doi: 10.1145/1349822.1349871

[139] Dag Sverre Syrdal, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2009. The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. In Proceedings new frontiers in human-robot interaction, a symposium at the AISB 2009 convention (pp. 1–8). Heriot Watt University, 8-9 April 2009, Edinburgh, Scotland.

[140] Xiang Zhi Tan, Samantha Reig, Elizabeth J. Carter, and Aaron Steinfeld. 2019. From one to another: how robot-robot interaction affects users' perceptions following a transition between robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 114-122. doi: 10.1109/HRI.2019.8673304

[141] Xiang Zhi Tan, Marynel Vázquez, Elizabeth J. Carter, Cecilia G. Morales, and Aaron Steinfeld. 2018. Inducing Bystander Interventions During Robot Abuse with Social Mechanisms. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (*HRI '18*). Association for Computing Machinery, New York, NY, USA, 169–177. DOI:10.1145/3171221.3171247

[142] Jonas Togler, Fabian Hemmert, and Reto Wettach. Living interfaces: The thrifty faucet. In Proceedings of the 3rd International Conference on Tangible and Embedded Interaction, pages 43{44. ACM, 2009. ISBN 978-1-60558-493-5. doi:10.1145/1517664.1517680.

[143] Deborah Tollefsen. 2005. Let's pretend! Children and joint action. Philosophy of the Social Sciences, 35, 1, 75-97.

[144] Michael Tomasello and Malinda Carpenter. 2007. Shared intentionality. Developmental science 10, 1, 121-125.

[145] Cristen Torrey, Susan R. Fussell, and Sara Kiesler. 2013. How a robot should give advice. In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 275-282. doi: 10.1109/HRI.2013.6483599.

[146] Eric Topol. 2019. Deep Medicine: How artificial intelligence can make healthcare human again. Basic Books. New York.

[147] Steven Umbrello. 2020. Meaningful human control over smart home systems: a value sensitive design approach. 40-65.

[148] Rik van den Brule, Gijsbert Bijlstra, Ron Dotsch, Pim Haselager, and Daniël H. J. Wigboldus. 2016. Warning signals for poor performance improve human-robot interaction. *J. Hum.-Robot Interact.* 5, 2 (September 2016), 69–89. DOI:https://doi-org.tudelft.idm.oclc.org/10.5898/JHRI.5.2.Van_den_Brule

[149] Jeroen van den Hoven. 1998. Moral responsibility, public office and information technology. In *Public Administration in an Information Age: A Handbook*, eds I. T. M. Snellen and W. B. H. J. Van de Donk (IOS Press), 579. Available at: https://books.google.nl/books?hl=nl&lr=&id=ehjvAgAAQBAJ&oi=fnd&pg=PA97&ots=JSy2OH4ng0&sig=jiK842VhBsC_h9oHmsY2YGufgeQ&redirесс=y#v=onepage&q&f=false

[150] J. J. Van der Waa, J. van Diggelen, M. Neerincx. 2018. The design and validation of an intuitive certainty measure. IUI 2018workshop on explainable smart systems. In: IUI 2018 workshop on explainable smart systems. ACM.

[151] Manuela M. Veloso. 2018. The Increasingly Fascinating Opportunity for Human-Robot-AI Interaction: The CoBot Mobile Service Robots. *J. Hum.-Robot Interact.* 7, 1, Article 5 (May 2018), 2 pages. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3209541

[152] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (*CHI EA '20*). Association for Computing Machinery, New York, NY, USA, 1–6. DOI:https://doi-org.tudelft.idm.oclc.org/10.1145/3334480.3381069

[153] Astrid Weiss, Nicole Mirnig, Ulrike Bruckenberger, Ewald Strasser, Manfred Tscheligi, Barbara Kühnlenz, Dirk Wollherr, and Bartlomiej Stanczyk. 2015. The interactive urban robot: user-centered development and final field trial of a direction requesting robot. J. of Behavioral Robotics, 6, 1. DOI:10.1515/pjbr-2015-0005

[154] Katherine S. Welfare, Matthew R. Hallowell, Julie A. Shah, and Laurel D. Riek. 2019. Consider the human work experience when integrating robotics in the workplace. In 14th ACM/IEEE international conference on human-robot interaction (HRI), IEEE, 75-84. doi: 10.1109/HRI.2019.8673139

[155] Julika Welge and Marc Hassenzahl. 2016. Better Than Human: About the Psychological Superpowers of Robots. In: Agah A., Cabibihan JJ., Howard A., Salichs M., He H. (eds) Social Robotics. ICSR 2016. Lecture Notes in Computer Science, vol 9979. Springer, Cham. https://doi.org/10.1007/978-3-319-47437-3_97

[156] David Whitney, Miles Eldon, John Oberlin, and Stefanie Tellex. 2016. Interpreting multimodal referring expressions in real time. In IEEE International Conference on Robotics and Automation (ICRA), IEEE, 3331-3338. doi: 10.1109/ICRA.2016.7487507.

[157] Tom Williams, Priscilla Briggs, and Matthias Scheutz. 2015. Covert robot-robot communication: human perceptions and implications for human-robot interaction. *J. Hum.-Robot Interact.* 4, 2 (September 2015), 24–49. DOI:https://doi-org.tudelft.idm.oclc.org/10.5898/JHRI.4.2.Williams

[158] H. James Wilson and Paul R. Daugherty. 2018. Collaborative intelligence: humans and AI are joining forces. Harvard Business Review 96, 4, 114-123.

[159] Heather Wiltse. 2020. Revealing relations of fluid assemblages. Relating to Things: Design, Technology and the Artificial, 239-254.

[160] Michael Wooldridge,and Nicholas R. Jennings. 1995. Intelligent agents: Theory and practice. The knowledge engineering review, 10, 2, 115-152.

[161] Robert H. Wortham and Vivienne Rogers. 2017. The muttering robot: improving robot transparency though vocalisation of reactive plan execution." In *26th IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man) Workshop on Agent Transparency for Human-Autonomy Teaming Effectiveness (Lisbon:)*. 2017.

[162] Yuto Yamaji, Taisuke Miyake, Yuta Yoshiike, P. Ravindra S. De Silva, and Michio Okada. 2011. Stb: Child-dependent sociable trash box. *Int. J. of Soc. Rob.,* 3, 4, 359-370. DOI:10.1007/s12369-011-0114-y

[163] Rayoung Yang and Mark W. Newman. 2013. Learning from a learning thermostat: lessons for intelligent systems for the home. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (*UbiComp '13*). Association for Computing Machinery, New York, NY, USA, 93–102. DOI:https://doi.org/10.1145/2493432.2493489

[164] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, IEEE, 1-8. doi: 10.1109/CIG.2018.8490433