Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

From Tools Towards Cooperative Assistants

Matti Krüger, Christiane Wiebel, Heiko Wersing

2017

Preprint:

This is an accepted article published in HAI 2017, 5th International Conference on Human-Agent Interaction. The final authenticated version is available online at: https://doi.org/[DOI not available]

From Tools Towards Cooperative Assistants

Matti Krüger **Christiane B. Wiebel** **Heiko Wersing**

Honda Research Institute Europe
Carl-Legien-Strasse 30
63073 Offenbach/Main, Germany
{matti.krueger,christiane.wiebel,heiko.wersing}@honda-ri.de

ABSTRACT

Endowing assistant systems with more autonomy establishes the transition from a human-controlled tool towards a self-directed agent capable of own decisions and goals. In this concept paper we suggest to perform the design of such an assistant agent according to principles of cooperativity. We first review definitions of cooperation between animals, humans and machines and then discuss advantages of cooperation also for a human-machine interaction system. We concentrate on the important roles of adaptivity and responsibility within the interaction. We argue that main benefits of a cooperative design are alleviation of typical automation issues like controllability, complacency, trust, and greater flexibility of the combined human-machine system in tasks with high variability.

ACM Classification Keywords

H.1.2 User/Machine Systems: Human factors; K.4.3 Organizational Impacts: Automation, Computer-supported collaborative work; H.5.3 Group and Organization Interfaces: Computer-supported cooperative work

Author Keywords

cooperation; human-machine interaction; adaptivity; assistant systems; autonomy

INTRODUCTION

There is a growing interest in designing assistant systems with more and more autonomy, driven from the demands of challenging applications like autonomous driving or home robots. Greater autonomy allows a person to delegate complete subtasks to an assistant system and reduce the human workload considerably. Another possible advantage is self-controlled system adaptation, enabling greater flexibility in highly variable scenarios. On the other hand there are prevalent negative automation side-effects affecting the human user, which are well known from ergonomics research: loss of expertise, complacency, controllability and trust issues [23, 10, 5]. We believe that in many situations the best way to approach these problems is by making the step from considering an assistant system as a tool towards designing it as a cooperation partner. The particular importance of mutual attention- and intention modeling for such a cooperative approach has been discussed by Wachsmuth [31] and demonstrated for a virtual embodied assistant [15]. Along this line we want to identify the qualitative steps with respect to mutual adaptivity and responsibility which are necessary for this cooperative transition and discuss advantages and possible disadvantages of this approach.

A good example for this desired change process is the concept of shared autonomy which has been introduced as a model of distributing control between human and machine in teleoperation scenarios [8]. Schilling et al. [27] recently proposed to extend the scope of shared autonomy from mere teleoperation of an intelligent tool towards more flexible human machine cooperation scenarios. They focus on the role of communication for negotiating the relation of autonomy and freedom on hierarchical levels separated into intentions, strategies and selection of means. The advantage of adaptivity in shared autonomy has recently been investigated by Nikolaidis et al. [19] for a multiple target robot grasping scenario. They showed that the co-adaptation of robot and human actions can deliver an optimal human-robot system performance with respect to the tradeoff between human trust and robot action effectivity.

In this manuscript we first review general definitions of cooperation and order them according to increasing requirements and shifts in the distribution of responsibilities for the cooperation partners and their interaction. We then discuss the special case of human-machine interaction in this context. In the subsequent section we define the transition stages from tools and adaptive tools towards cooperative assistants up to long-term and sustained cooperation. We conclude with an overview summary of our argumentation.

COOPERATION DEFINITIONS

Defining and explaining cooperative behavior has been an important research topic in different disciplines including philosophy, biology, sociology, psychology and economics (e.g. [24, 1, 20, 17]), trying to answer questions like: *What is cooperation?*, *How is cooperative behavior initialized and maintained?*, *What are requirements and benefits of cooperative behavior?* and *How is behavior shaped by the intent to cooperate?* We review some prominent definitions with increasing requirements in the following:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HAI 2017, October 17–20, 2017, Bielefeld, Germany.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5113-3/17/10 ...$15.00.
http://dx.doi.org/10.1145/3125739.3125753

The Oxford dictionary of current English gives a universal definition of the act to cooperate as follows:

> "work or act together in order to bring about a result"

This broad definition highlights two fundamental aspects of cooperation: (1) cooperation involves a plurality of agents and (2) for the purpose of a certain result they will coordinate their actions. Piaget's 1965 definition of cooperation goes one step further. He states (translated by Hoc [11]):

> "Cooperating in action is operating in common, that is adjusting the operations performed by each partner by the means of new operations. It is coordinating each partner's operation into a single operation system, the collaboration acts themselves constituting the integrative operations."

Importantly, Piaget's definition makes it explicit how two cooperating partners act together - by means of adjusting their behavior. Even more specific is the following definition by Hoc [11]. He states:

> "Two agents are in a cooperative situation if they meet two minimal conditions. (1) each one strives towards goals and can interfere with the other on goals, resources, procedures etc. (2) each one tries to manage the interference to facilitate the individual activities and/or the common task when it exists. The symmetric nature of this definition can be only partly satisfied."

Hoc [11] considers cooperation as a subclass of collective activity which should be described as *interference management in real-time*. That is, in addition to Piaget's definition, he introduces a temporal dimension to his definition. The regulative process has to happen in real-time excluding the case that the agents' relationship and interactions have been defined entirely by a designer or manager in advance. Besides, he differentiates the purpose of the cooperation into a common goal or each agent's personal matter. In his definition of cooperation, both concepts of the motivation to cooperate would be valid. To what extent the agents act on behalf of their own interest or a common goal can also be further characterized by their support for each other. Bratman [6] adds another defining component in his definition of a *shared cooperative activity*. In addition to mutual responsiveness and commitment to the joint activity, he suggests the concept of *mutual support*. This means the willingness of all agents to support one another in performing their role in the cooperative activity. To summarize, critical differentiations between definitions of cooperation can be the degree to which each agent in a cooperative system works towards the same or different goals, to what degree the cooperative behavior is volitional and self-driven or whether the cooperation partners are mutually supportive.

Cooperation Complexity

Different levels of complexity in cooperative behavior have been defined. Boesch et al. [3] proposed a categorization of levels of cooperative behavior in chimpanzees performing group hunt. They differentiate between similarity hunt, coordination hunt, synchrony hunt and collaborative hunt. Similarity hunt means that hunters perform similar actions towards the common goal of hunting prey but without any coordination in space and time. In synchrony hunt, hunters adjust their actions in time but not in space. They may for instance begin at the same time or adjust their speed to each other. The next level, coordination hunt, includes the dimension of space. The highest level of hunting behavior is described in collaboration hunt. Here, hunters perform different and complementary actions in order to hunt the same prey. Brinck et al. [7] argue that future-oriented cooperative behavior in addition has to involve shared representations that go beyond things that are immediately present. According to the authors, this means that agents need to be able to communicate about *things that do not yet exist*. This suggests an important role of symbolic communication for future-oriented cooperative behavior.

Representation Sharing

In order to have cooperation in a narrow sense, as for example suggested by Hoc or Bratman [11, 6], the involved agents need internal models or representations that allow them to anticipate future states of themselves, the other agent and the task. Moreover, the ability to communicate these representations seems crucial for the cooperation success. The most basic part of such communication corresponds to the ability to direct someone's attention to the own focus of attention [28]. Humans for instance use mainly deictic gaze or pointing gestures to do so and are extremely accurate in their interpretation (for a review see.: [29]). Joint attention can thus provide the basis for a *common perceptual ground* [28]. It has also been shown for human-robot interaction that joint attention is a crucial mechanism for successful joint actions, leading to reduced task time and less errors [18]. It should be noted that shared representations between agents in a cooperative system have been termed differently by different authors (e.g.: [11, 14]). The overall idea however remains the same. Shared representations or notions of internal models as well as the ability to communicate or at least direct the attention of another agent within a joint reference framework are undoubtedly critical features for cooperative behavior. Furthermore, the degree to which such representations exist and are shared represents a critical criterion for distinguishing between levels of cooperation.

COOPERATION IN HUMAN MACHINE INTERACTION

In the context of human-machine interaction, the concept of cooperation has gained increasing interest along with the rapidly growing capabilities of autonomous or partly autonomous systems (e.g.: [9, 2, 11, 13]). A key question that has been discussed is how to differentiate human-machine cooperation from basic interaction.

Adaptability

Hollnagel at al. [14] were the first to characterize human-machine interaction as an interaction between two cognitive systems. Until then, human-machine interaction had been primarily concerned with the physical adaptation of the human and the machine. This meant mostly the adaptability of a machine's design to the physical condition of a human operator (e.g.: Make a system adjustable so that a wide range of operators can physically reach the emergency button). By

means of this change in perspective Hollnagel et al. [14] established the term *cognitive system engineering*. Although the authors did not actively use the term cooperation, much of what they claimed can be considered as a cooperative approach for human-machine interaction. They elaborated that for further progress in the field, *cognitive adaptability*[1] had to be added to the system. They declared that besides physical characteristics, cognitive characteristics of the human user would have to be acknowledged in the design process of human machine systems. They pointed to the necessity of *"dynamically adaptable"* internal models for the user and the machine and their interactions. Furthermore, they noted that this has to be considered with caution, as the physical and the psychological world can be quite different. Physical quantities might not translate one-to-one into psychological entities in which case their correspondence has to be worked out (e.g. [26]).

More recently, Hoc [13] formulated his idea of a human-machine cooperation framework as opposed to human-machine interaction, in line with the ideas formulated by Hollnagel et al. [14]. Two aspects in particular were considered in his work. The need for uncertainty and risk management as a result of the dynamic nature of tasks which are not fully scripted beforehand and an appropriate and dynamic function allocation between human and machine that takes into account the peculiarities of a certain situation. Both points imply that the whole system consisting of human and machine is able to adapt itself continuously.

Observability and Directability

Christofferson et al. [9] have argued that in order to turn a machine into a team player the *cooperation automation* has to be observable and directable. Once again, observability relates to the idea of shared representations or transparency. The authors mainly stress the necessity of the human user to understand the machine's state but they do not take into account that the machine must also understand the human to some extent. This seems however equally important for a successful interaction. Directability on the other hand picks up the topic of shared authority. The authors state that as long as responsibility has to remain with the human, authority is not questionable. The human has to be able to influence the machine's activities. Therefore, directability is meant only in one direction (human-machine) here. Related concepts are also discussed by Bengler et al. [2] who separate key elements of human-machine cooperation into five layers: i) intention, ii) modes of cooperation suggested by [12], iii) allocation as a mediator between interaction and interface, iv) interface itself and v) physical contact layer (possibly limited or forbidden).

INTERACTION LEVELS

Besides having the ability to extend our operation space through tool use, the large extent of cooperation within human societies is thought to be a major reason for human evolutionary success [4].

[1]We refer to adaptability as the possibility for an entity to have parameters of itself changed by another entity whereas adaptivity describes the ability of an entity to change parameters of itself. Hollnagel et al. do not specify their understanding of *adaptibility* but appear to use the term in the sense of our understanding of adaptivity.

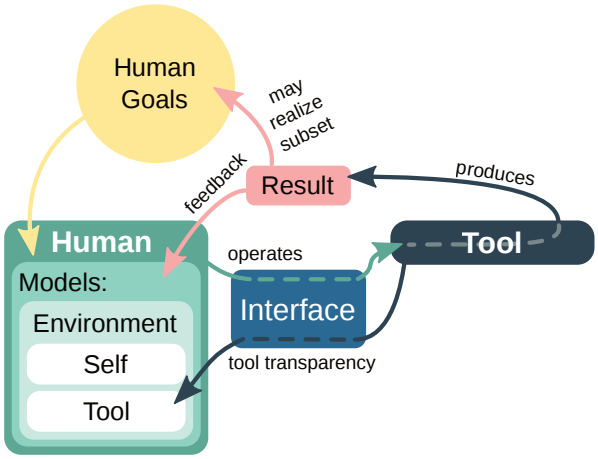

Figure 1. Interaction during (non-adaptive) tool use.
A tool extends the human capabilities to achieve a result. Its successful operation relies on a proper human understanding of the tool's properties and features, accessible via an interface that matches the human sensory modalities. Results produced by tool use may also be used as feedback for better tool-understanding and improvements in future success. The human user carries full responsibility for adapting the usage of the non-adaptive tool to variable task conditions.

Here we describe how the shift from tool use to cooperation can be characterized by differences in the distribution of responsibilities among agents. Thereby we illustrate how changes in this distribution depend on varying requirements for adaptive abilities.

We focus on three main stages of interaction that take the steps from tools to cooperative assistants. Finally, we discuss additional mechanisms that aid in providing stability to interactions between highly adaptive agents. By defining and illustrating these stages we aim to support the classification of human-machine systems as well as the identification of entailed requirements and consequences.

Tool Use

We define a tool as an entity that is used in order to extend a person's capabilities and efficiency in carrying out a task. We consider here only non-adaptive tools which means that all responsibility for adapting the tool usage pattern in flexible task settings rests with the user. For purposeful use, the user needs to be able to access some of the tool's properties via an interface (Figure 1) which allows him or her to develop an understanding (model) of the tool. These properties thereby need to be accessible in modalities that match the user's available sensors. Besides having a model of the tool, a user also needs to understand the tool in relation to itself and possibly other relevant components of the environment. Results produced by tool use may be used as feedback for better tool understanding and improvements in future success.

We include in our non-adaptive tool definition also complex machines like automated robots, as long as they are employed by a human to achieve a well-defined goal and are not capable of self-adaptation according to changing environmental conditions.

Example - Hammer

A prototypical example for a non-adaptive tool is a hammer. The geometry and material composition of a hammer allow its user to carry out the task of driving nails into other materials. Note, however that the hammer obtains its purpose only through use in a task by an external agent. If someone exploits the hammer's properties to use it to prevent papers from flying from a desk, its purpose changes from an object for hammering nails into materials into a hammer-shaped paperweight. In either case, the hammer itself carries no direct responsibility in its actions because it relies on responses made by others in order to be involved in executing a task. The lack of self-induced change paired with a complete reliance on user involvement makes it relatively easy for a user to understand and predict the effects of hammer-use. However, it also means that a simple hammer is unable to perform sub-tasks independent from a user and is therefore also limited in utility by its user's abilities[2].

Example - Robot Vacuum Cleaner

Robot vacuum cleaners are well-established automatic home robots with a defined behavior and clear purpose. Let us explain in the following under which conditions we consider them still an example for a tool rather than a cooperative assistant.

A robot vacuum cleaner is capable of moving along a floor and vacuuming it in an automatic manner. It contains sensors for detecting collisions with other objects and is able to react to such collisions with a predefined maneuver such as briefly moving backwards and turning by some random angle. With these abilities, the robot can act autonomously within a restricted operation space. We assume that there is no persistent memory for building a map of the environment which is also not the case for many current robot vacuum cleaners.

The model which underlies its actions produces flexible reactive behavior. However, the model itself is not flexible. When the vacuum cleaner bumps into a wall, it will do so again every time it approaches that wall from a similar initial position. Similarly, if it gets stuck in a carpet once, it will repeat this mistake whenever encountering the carpet, rather than trying to avoid the carpet after the first failure. The robot must therefore rely on the human user to make adaptations in order to provide an environment that is suitable for the vacuum cleaner to operate without errors. Its user carries the responsibility for allocating a suitable operation space and if necessary make appropriate adaptations to ensure tool functionality.

Consequences

When using a tool, the full responsibility resides with the human user in the interaction. For a simple tool this may offer a clear advantage by extending the user's abilities combined with good controllability. For a complex tool, which may even be capable of automatic reactive behavior, adapting the tool usage pattern and environment may become a high burden for the human user. Consequently, means for self-adaptation of the tool may become necessary. We will discuss this in the following.

[2]A cognitive bias known as *law of the instrument* [16, p.15] or *Maslow's hammer* is exemplified by the saying "if all you have is a hammer, everything looks like a nail". It refers to the observation that people often use tools which they are familiar with beyond the tool's intended scope which may be inappropriate when considering available alternatives. Consequently human adaptation to a tool may also limit the human flexibility of adapting to different task settings.

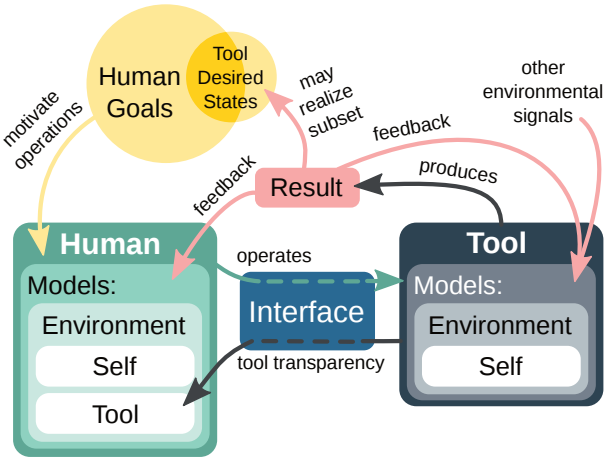

Figure 2. Interaction with adaptive tools.
Using experience, the adaptive tool maintains a model of the environment and itself to adapt its parameters for improved operation towards a goal. This can reduce the adaptation load for the human user. Conflicts can occur if tool and human goals mismatch.

Adaptive Tool Use

An adaptive tool (see Figure 2) is a tool which contains adaptive components, i.e. which has the ability to change one or more of its own parameters in response to environmental variations. In addition to a tool capable of producing flexible behavior, in an adaptive tool the model that underlies its operations is flexible as well. It thus has the ability to learn from experience. Consequently, in the ideal case it can reduce the adaptation load that has to be carried by the human interaction partner in flexible task situations.

Example - Cruise control

One example for an adaptive tool is the basic cruise control function, also known as speed control or tempomat which is available in many motor vehicles. It is characterized by an automatic control of the acceleration of a motor vehicle to maintain a speed value defined by its user. When active, cruise control takes over the operator's task of controlling vehicle acceleration. However, it does so in a way which only permits the realization of its desired speed value, independent of how sensible maintaining this particular speed is in a given situation.

Therefore the user has to terminate or manually adapt the cruise control functionality whenever the maintenance of a certain speed clashes with more important goals such as safety and complying with traffic rules. Because the cruise control function is substitutive by nature, it can furthermore have negative effects on a user's ability to control vehicle speed and react appropriately upon takeover of responsibility after function termination [32].

Example - Adaptive Robot Vacuum Cleaner
We take again the example of a home robot vacuum cleaner, but extend it by the capability to modify its operation model based on its operation history. Using more advanced sensing it could e.g perform self-localization and map building. Based on its position estimation information it could try to avoid frequent bumps into obstacles or adjust time and effort to improve overall cleaning efficiency. It is, however, not guaranteed that this self-driven optimization coincides with the desired cleaning pattern of the user. To achieve this, a means for interactive negotiation of proper cleaning targets which are also accessible to the robot within the flat would be necessary.

Consequences
Adaptivity can lead to a self-controlled flexibility for environmental variations and thus not only allow for wider compatibility compared to static entities but also for a delegation of responsibility to the adaptive entity. On the contrary, adaptive tools may follow their own goals and constraints that make it harder to use an adaptive tool in a way that goes beyond the intention of its designer. Such limitations typically arise when the models that underlie adaptive operations substitute more flexible models available to an operator. Especially in human-machine interaction with partially automated components that could also be performed by humans this can frequently be the case. A good example for challenges of user interaction with adaptive tool-like assistants is the interaction with the *NEST* learning thermostat, which has the ability to learn a typical climate and heating schedule based on the history of manual user settings during day and week. In a user study Yang et al. [33] showed that users of this system found the intelligent learning functions hard to understand and the resulting schedules were often not matching the original user intent. Resolving this issue would require better cooperative means for establishing a shared goal between user and system.

Besides substitution-based restrictions, increased automation can introduce various secondary issues and challenges concerning the role of a human user that have been frequently discussed [23, 10, 5]. Parasuraman and Riley [23] identified such problems in a systematic manner by means of analyzing for instance failures and accidents in aviation. The authors described four main types of difficulties in the interaction of humans with automation: loss of expertise, complacency, trust and loss of adaptivity. Loss of expertise for instance, describes the circumstance under which a human operator becomes passive and less vigilant in the presence of a machine that fulfills an autonomous role. As a consequence, taking over responsibility in case of shortcomings of the machine or unpredicted situations might result in delayed or poor reactions by the human operator which may impose a substantial danger in many scenarios.

Cooperative Assistance
Let us first summarize the requirements for cooperative human-machine interaction aggregated in the previous sections:

Cooperation occurs between agents if they adapt to the state and actions of the other agent in a manner that facilitates the realization of a shared cooperation goal. This adaptation requires mutual models and understanding with respect to

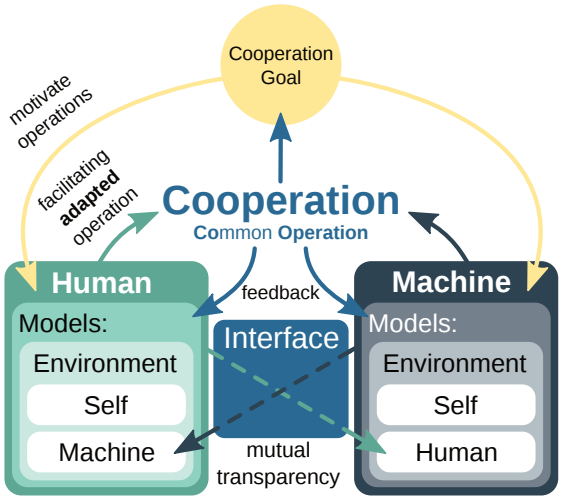

Figure 3. Interaction with a cooperative assistant.
In human machine cooperation the machine is capable of relating the states and operations of a human agent to the realization of a cooperation goal and can adapt its own states and operations in a way that, in conjunction with the human operations, facilitates goal realization. The interface allows a mutual adaptation by maintaining proper models of the cooperation partner. The changing responsibilities are negotiated via the interface and adapted according to the flexible task setting.

the intentions, actions and plans that are relevant for the goal realization. The development and maintenance of such models requires mutual transparency and communication of relevant variables by the cooperating agents. Cooperative assistance is then the application of the cooperative human-machine interaction principles to a human-supporting assistant system (see Figure 3).

A principle motivation for cooperation is to facilitate and potentially enhance the actions and capabilities of each agent alone by dynamically merging their available resources. This sharing of resources goes beyond a substitution approach in that it does not merely aim for a creation and an exclusive use of functional redundancies but rather a context-dependent distribution of available resources that allows the agents to tackle a larger set of tasks. Such distributions should furthermore not be limited to scenarios in which cooperating agents may act independently but also take advantage of capabilities that arise or at least benefit from interdependent use of resources [5].

Example - Cooperative Robot Vacuum Cleaner
An important addition which a cooperative assistant needs in comparison to an adaptive tool is the ability to model the behavior of another agent in relation to a goal as well as its own actions and abilities. This can be summarized as a goal-oriented adaptivity. In addition a cooperative assistant should be able to

- acquire information about a human cooperation partner that is potentially task relevant

- relate this information to a cooperation goal and its own actions and abilities in order to produce purposefully adapted behavior

- provide information about its own states and actions to the human cooperation partner in order to allow for mutually adapted operations

Furthermore, a cooperation goal which motivates the actions of human and robot is needed to direct adaptivity. In the case of the cleaning robot example this could for instance be the (high level) goal of keeping an apartment clean.

To better achieve this goal, on a basic level, the robot could start by recognizing areas which a human is taking care of and plan its own actions to minimize overlap with these areas, thus promoting efficiency. It could also learn from human behavior which places and actions are common sources of dirt and focus on these in particular by for instance more frequent coverage or even by informing the human cooperation partner about its observations in hope of more considerate future behavior. Besides, it could notify its human cooperation partner about which places are difficult or impossible for it to clean in order to let the partner more easily determine where to best apply his or her resources. This could for example be narrow corner regions or areas to which access is blocked by an object the robot can't or isn't authorized to move. The human partner could then support the robot by cleaning the difficult areas or moving obstacles out of the robot's way respectively. After moving an object out of the way, the human could in turn tell the robot that the problem has been taken care of upon which the robot could expand its planned operation space. Additionally the robot could react to human instructions on what areas to take care of and about what types of issues it should request human support.

Note that although the cooperative approach imposes additional requirements on the robot's communicative and modeling abilities, it may also help to keep requirements low in other domains without sacrificing overall effectiveness. In particular, it facilitates a dynamic and purposeful distribution of available resources such that for example a cleaner robot cooperating with a human won't need the ability for cleaning narrow spaces which the human partner could take care of with little effort.

Consequences

Cooperating agents need to be able to perceive information to infer their partners' goals and adapt their actions appropriately. This property enables a dynamic negotiation of individual responsibilities with respect to each individual's goals and thus more robust and sustainable task success. This might especially be true if the nature of the task or an individual agent's capabilities may vary over time or across situations. Failures or disengagement of one of the agents might be more easily detected and confronted within the system.

Such personalized adaptivity however, also has a potential for introducing novel cases of interaction failure and conflict due to model complexity and a risk of model mismatch. Human factors with potential for personalized adaptivity can sometimes be highly abstract concepts that are not expressed in a universally applicable and unambiguous manner. For instance, behavioral indicators for satisfaction with a function may not only differ across different people but even within individuals in different situations and even spoken language can often only be disambiguated in a larger context and can contain dangerous pitfalls such as irony. For a narrow set of scenarios, simplistic models that are easy to understand for a user could suffice as a basis for personalized adaptivity. But to account for variability in the expressions of such concepts, the respective models may need higher complexity as well. With increasing complexity the difficulty for a user to correctly understand the adaptive process increases as well. To promote successful interaction with assistive tools, efforts towards making models transparent to a user may have to increase with the respective model's complexity.

The requirement for transparency of personal variables means that some sharing of personal information is necessary. In many cases this information may be taken from variables that are often somewhat publicly accessible such as the visual appearance. However, some scenarios can rely on the exposure of information that is considered to be private. In such cases a trade-off between privacy and utility arises and the notion of trust gains importance besides its possible role in immediate interaction success: Appropriate trust in the functionality of a cooperation partner should be present to promote successful adaptive interactions. Furthermore, appropriate trust in data confidentiality allows individuals to judge whether the gain through cooperation surpasses the cost entailed in sharing private information. Unless private information is not sufficiently valued by involved individuals, confidentiality may thus become a decisive factor for the success of cooperation and should accordingly be considered with care.

In cooperation no agent is assumed to have an idle or more relaxed role but is supposed to actively contribute to synergy-enabling resource distributions. As a result, typical problems of automated functions like being "out-of-the-loop", loss of expertise or complacency [23, 10, 5] should less likely occur. A good distribution of resources may also reduce the risk of individual skill loss. If a cooperative machine would, as part of the cooperative action, take over a task that would otherwise be performed by a human operator, the skills of the operator may still suffer with respect to that task. However, this loss can happen in exchange for an improvement in other tasks or components of tasks that are relevant in the cooperative work.

Towards Sustainable Cooperation

By sustainable cooperation we denote an extension of basic cooperation that is targeted at establishing long-term cooperative success between multiple agents by introducing requirements that should facilitate the formation of mutually beneficial long-term and interdependent relationships.

Establishing long-term stability faces two main challenges: i) Robustness with regard to cooperation failures and ii) the necessity to accept short-term disadvantages for advantages that only pay-off in the long-term collaboration.

Let us first focus on the issue of cooperation failures. In human-machine interaction, inappropriate trust in the abilities of the partner is a potential source of cooperation failure and inefficiency: Without trust, a human is unlikely to engage in cooperative activity with a machine. When the trust is inap-

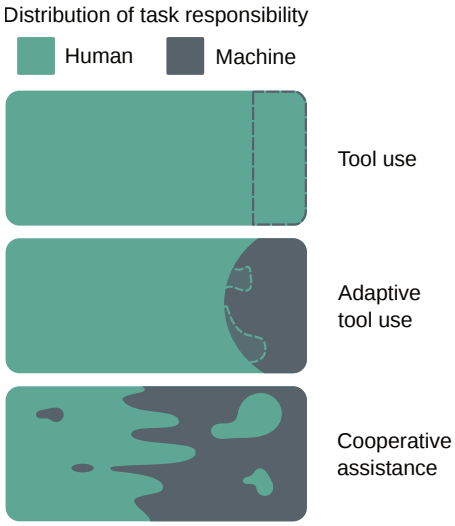

Figure 4. Distribution of task responsibilities for different levels of human-machine interaction.
Using a tool extends the accessible task space (represented by the area surrounded by a dashed grey line). Still, the full responsibility and adaptation load is carried by the human tool user. With an adaptive tool, part of the task responsibility is taken over by the adaptive tool. However, some of the operations carried out by the adaptive tool may clash with the goals of its user who would perform differently if this responsibility was allocated to him or her (represented by the grey-green areas surrounded by dashed lines). With cooperative assistance, the distribution of task responsibilities is flexible and can be negotiated between human and machine due to goal-directed mutual adaptivity. Imbalances as described for the adaptive tool can be avoided by allocating responsibilities according to competencies and momentarily available resources. Furthermore, flexible responsibility negotiation may counteract occurrence of automation problems.

propriately high, an agent may neglect its role in monitoring and interfering with another agent's actions. When trust is inappropriately low, an agent may waste available resources on monitoring and interfering with another agent's actions and thus potential of the cooperating system remains unused.

Trust is also relevant for cooperation success in the sense of having trust in benefiting from the interaction. Consequently we regard *social tolerance* [25], the ability of an agent to accept that cooperating agents must all benefit from the cooperation outcome, as an important prerequisite for sustainable cooperation. Implementing prosocial behavior means that substantial effort may have to be made towards ensuring that a cooperating agent's goals are correctly identified and put in relation with one's own as well as other agents' actions. On a similar note, the application of strategies which use a more sophisticated concept of reward may help to sustain cooperative behavior. Specifically, three R's [4], i.e. reputation, reciprocity and retribution have been argued to explain the development of stable cooperation within cultural groups [30, 21, 22] and could aid in designing reward functions of artificial agents.

Examples for short-term disadvantages that need to be overcome for sustained cooperation are i) human learning effort in understanding the machine assistant and its capabilities, ii) inefficient transition times that are needed for the convergence towards an efficient stable cooperation pattern, and iii) effort for establishment of a common long-term goal beyond short-term goals.

One of the main benefits of sustained long-term assistive cooperation is the possibility to follow long-term goals like maintaining a proper balance between offering assistance and motivating the human to stay active for avoiding a loss of expertise and proficiency. This requires proper social cues in the interaction and can pave the way towards a real partner relation between human and machine [31].

CONCLUSION

We have discussed requirements for elevating an assistant system from a simple tool towards a cooperation partner. Our main argument is that increasing autonomy of the assistant should be accompanied by increasing adaptivity towards the human partner. This progressive change can consequently enable a redistribution of responsibilities in an interaction and we argue that more complex task settings require a more flexible distribution of responsibility between human and system (see Figure 4). We consider the communicative processes for establishing cooperation as a necessary prerequisite for an optimal distribution of responsibility with regard to criteria of trust, mutual support and transparency.

REFERENCES

1. R Axelrod and WD Hamilton. 1981. The evolution of cooperation. *Science* 211, 4489 (1981), 1390–1396.
2. Klaus Bengler, Markus Zimmermann, Dino Bortot, Martin Kienle, and Daniel Damböck. 2012. Interaction principles for cooperative human-machine systems. *it-Information Technology* 54, 4 (2012), 157–164.
3. Christophe Boesch and Hedwige Boesch. 1989. Hunting behavior of wild chimpanzees in the Tai National Park. *American journal of physical anthropology* 78, 4 (1989), 547–573.
4. Robert Boyd and Peter J Richerson. 2009. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1533 (2009), 3281–3288.
5. Jeffrey M. Bradshaw, Robert R. Hoffman, Matthew Johnson, and David D. Woods. 2013. The Seven Deadly Myths of "Autonomous Systems". *IEEE Intelligent Systems* 27, 2 (2013), 43–51.
6. Michael E Bratman. 1992. Shared cooperative activity. *The philosophical review* 101, 2 (1992), 327–341.
7. Ingar Brinck and Peter Gärdenfors. 2003. Co–operation and communication in apes and humans. *Mind & Language* 18, 5 (2003), 484–501.
8. Bernhard Brunner, Gerd Hirzinger, Klaus Landzettel, and Johann Heindl. 1993. Multisensory shared autonomy and tele-sensor-programming-key issues in the space robot technology experiment ROTEX. In *Intelligent Robots and Systems' 93, IROS'93. Proceedings of the 1993 IEEE/RSJ International Conference on*, Vol. 3. IEEE, 2123–2139.
9. Klaus Christoffersen and David D Woods. 2002. How to make automated systems team players. *Advances in human performance and cognitive engineering research* 2 (2002), 1–12.

10. Jean-Michel Hoc. 2000. From human - machine interaction to human - machine cooperation. *Ergonomics* 43, 7 (2000), 833–843.
11. Jean-Michel Hoc. 2001. Towards a cognitive approach to human–machine cooperation in dynamic situations. *International journal of human-computer studies* 54, 4 (2001), 509–540.
12. Jean-Michel Hoc. 2007. Human and automation: a matter of cooperation.. In *HUMAN 07*, A. Pruski (Ed.). Université de Metz, Timimoun, Algeria, 277–285.
13. Jean-Michel Hoc and Serge Debernard. 2002. Respective demands of task and function allocation on human-machine co-operation design: a psychological approach. *Connection science* 14, 4 (2002), 283–295.
14. Erik Hollnagel and David D Woods. 1983. Cognitive systems engineering: New wine in new bottles. *International Journal of Man-Machine Studies* 18, 6 (1983), 583–600.
15. Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. 2005. Lecture Notes in Computer Science. Springer-Verlag, London, UK, UK, Chapter A Conversational Agent As Museum Guide: Design and Evaluation of a Real-world Application, 329–343.
16. Abraham H Maslow. 1966. The Psychology of Science: A Reconnaissance. (1966).
17. Henrike Moll and Michael Tomasello. 2007. Cooperation and human cognition: the Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 362, 1480 (2007), 639–648.
18. Bilge Mutlu, Allison Terrell, and Chien-Ming Huang. 2013. Coordination mechanisms in human-robot collaboration. In *Proc. of the Workshop on Collaborative Manipulation, 8th ACM/IEEE Int. Conf. on Human-Robot Interaction*.
19. Stefanos Nikolaidis, Yu Xiang Zhu, David Hsu, and Siddhartha Srinivasa. 2017. Human-Robot Mutual Adaptation in Shared Autonomy. *Proc. Conf. on Human Robot Interaction, Vienna* (2017).
20. Martin A Nowak. 2006. Five rules for the evolution of cooperation. *Science* 314, 5805 (2006), 1560–1563.
21. Martin A. Nowak and Karl Sigmund. 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393, 6685 (1998), 573–577.
22. Karthik Panchanathan and Robert Boyd. 2003. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of theoretical biology* 224, 1 (2003), 115–126.
23. Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 2 (1997), 230–253.
24. Jean Piaget. 1965. *Études sociologiques*. Vol. 32. Librairie Droz.
25. Friederike Range and Zsófia Virányi. 2015. Tracking the evolutionary origins of dog-human cooperation: the "Canine Cooperation Hypothesis". *Frontiers in Psychology* 5 (2015), 1582.
26. Sverker Runeson. 1977. On the possibility of "smart" perceptual mechanisms. *Scandinavian journal of psychology* 18, 1 (1977), 172–179.
27. Malte Schilling, Stefan Kopp, Sven Wachsmuth, Britta Wrede, Helge Ritter, Thomas Brox, Bernhard Nebel, and Wolfram Burgard. 2016. Towards A Multidimensional Perspective on Shared Autonomy. In *2016 AAAI Fall Symposium Series*.
28. Natalie Sebanz, Harold Bekkering, and Günther Knoblich. 2006. Joint action: bodies and minds moving together. *Trends in cognitive sciences* 10, 2 (2006), 70–76.
29. Stephen V Shepherd. 2010. Following gaze: gaze-following behavior as a window into social cognition. *Frontiers in integrative neuroscience* 4 (2010), 5.
30. Robert L Trivers. 1971. The evolution of reciprocal altruism. *The Quarterly review of biology* 46, 1 (1971), 35–57.
31. Ipke Wachsmuth. 2015. *Embodied Cooperative Systems: From Tool to Partnership*. Springer International Publishing, Cham, 63–79.
32. Lingyun Xiao and Feng Gao. 2010. A comprehensive review of the development of adaptive cruise control systems. *Vehicle System Dynamics* 48, 10 (2010), 1167–1192.
33. 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*. ACM, 93–102.