Bringing Together Ergonomic Concepts and Cognitive Mechanisms for Human-AI Agents Cooperation

Marin Le Guillou, Laurent Prévot, Bruno Berberian

SPECIAL ISSUE AI IN HCI

Bringing together ergonomic concepts and cognitive mechanisms for human - AI agents cooperation

Marin Le Guillou[a,b], Laurent Prevot[b] and Bruno Berberian[a]

[a]ONERA, The French Aerospace Lab, Information Processing and Systems Department, 13661 Salon Cedex Air, France; [b]Laboratoire Parole Language, Aix-Marseille Université, CNRS, Aix-en-Provence, France

ARTICLE HISTORY
Compiled September 12, 2022

ABSTRACT
The deployment of artificial intelligence from experimental settings to concrete applications implies to consider the social aspects of the environment and consequently to conceive the interaction between humans and computers endowed with the aim of being partners in action. This paper proposes a review of the research initiatives regarding human-artificial agents interaction, including eXplainable Artificial Intelligence (XAI) and HRI/HCI. We argue that even if vocabulary and approaches are different, the concepts converge on the necessity for the artificial agents to provide an accurate mental model of their behavior to the humans they are interacting with. This has different implications depending on whether we consider a tool/user interaction or a cooperation interaction - which is far less documented despite being at the heart of the future concepts of autonomous vehicles. From this observation, the paper uses the cognitive science corpus on joint-action to raise finer cognitive mechanisms proved to be essential for human joint-action which could be considered as cognitive requirements for future artificial agents, including shared task representation and mentalization. Finally, interactions content hypotheses are arisen to satisfy the identified mechanisms, including the ability for the artificial agent to elicit its intentions and to trigger mentalization toward them from the human cooperators.

KEYWORDS
Human-Machine Teaming; eXplainable Artificial Intelligence; Joint-action; Mental models;

1. Introduction

Academic technical progresses within AI techniques have led to outstanding capabilities demonstrations such as AlphaGo [Silver et al., 2016]. Consequently industrial companies have started to fund and realize projects integrating these techniques, from personal assistants to autonomous cars. Indeed the 2019 AI Index Report [Perrault et al., 2019] shows an average annual growth of 48% between 2010 and 2018 for AI private investments, reaching $40 billions in 2018. To our understanding the advent of AI offers a new sub-field to Human-Computer Interaction(HCI) by providing computers with *agents'* abilities, where an *agent* is "a person or entity that acts or has the capacity to act, particularly on behalf of another or of a group" (American Psychological Association, APA). Artificial agents (AAs) aim to integrate society at

CONTACT Marin Le Guillou. Email: marin.le_guillou@onera.fr

various levels by replacing humans in tasks they want to delegate (3Ds tasks for Dirty, Dangerous and Demeaning) or are not as good as AAs at (for example big data processing) but also - and especially - by taking part to a human-AA cooperative system where every stakeholders' skills are efficiently operated towards a common goal. As the advent of the new AI techniques has been mainly driven by neural networks - which huge potential comes at the price of opacity paving the way for trust and acceptability questions - the whole interaction of AAs with humans has to be conceived.

The prevailing approach within AI research regarding cooperating humans and AAs assumes that the combination of the growing powers of models and computation will eventually allow to integrate human partners as any other environment variable. The idea of modeling system's users as a whole and by this mean meeting their expectations about system interaction is not new, as the literature of the period of expert systems shows [Kass and Finin, 1988], [Carberry, 1988], [Finin and Drager, 1986]. Today this idea can be retrieved in the context of (deep) Reinforcement Learning (RL) agents [Sutton and Barto, 2018] which relative RL techniques have brought to AI its best successes, such as AlphaGo. In this research field many researchers advocate for an approach of AI techniques where the less prior environmental knowledge the system designer includes, the best it is. In this logic, time is at discovering the best RL architecture for including humans as a variable while designing agents for cooperative tasks. Despite research work on this topic severely lacking of user studies, recent efforts have to be noted. Indeed, [Carroll et al., 2020] proposes a testbed allowing to make user studies on an adapted version of the popular game "Overcooked". This game allows to measure a performance on a cooperative task requiring goal and spatial coordination. The paper's authors use this testbed to show evidences about their model and learning strategy to acquire the necessary skills to successfully perform the task with humans. The combination of AI techniques proposal using an user study for evaluation can also be found in [Choudhury et al., 2019], which tests different architectures of AAs being inspired by the knowledge in cognition.

Without presuming nor denying a future breakthrough within AI techniques allowing such human modeling, we argue that the knowledge about human-human cooperative interaction could at worst speed-up the advent of cooperative AAs and drive the AI model design. The approach considering the knowledge from social sciences which will be developed in this paper led to the introduction of the concept of eXplainable Artificial Intelligence (XAI), or how to allow AI algorithms to explain their own behavior. The effort put into this problem has made it possible to progress in our understanding of what a good explanation is in terms of acceptability, what makes a human trust or distrust an algorithm. On the contrary, today little is known about what a good explanation is in terms of cooperation - *a process whereby two or more individuals work together toward the attainment of a mutual goal or complementary goals (APA)* - even if the sophistication of these algorithms makes them more than simple tools, precisely real teammates. In this paper, we assume that a prerequisite to facilitate the adoption of AI tools and their integration into operations is the identification of the information to be provided to enable the human operator to work in cooperation with AI. The objective of this paper is to bring together the recent advances around the notion of explainability for the human operator, to pose its limits and to propose theoretical frameworks allowing to overcome them. In this sense, this paper will first depict the minimum interaction requirements for AAs with a trust and acceptability focus, before focusing on special abilities required for AAs sharing high-level objectives - implying real time task allocation and coordination.

2. The need for eXplainable Artificial Intelligence (XAI)

When considering the ability of a human being to coordinate with a partner (human or artificial), the development of an accurate mental model of that partner's behavior appears to be a central element. Rouse and Morris [Rouse and Morris, 1986] gave a widely accepted definition of Mental Models (MMs) as "mechanisms whereby humans are able to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future states". The accuracy of mental models have been known for leveraging both the performance in using "simple" tool systems [Kieras and Bovair, 1984] (in this case the mental model means "how it works" knowledge) and in team work [Lim and Klein, 2006] (where mental model is extended to *shared* mental model, itself including *team model* designating team members' understanding of each others). However, with increase in system complexity (for example, the multiplication of the number of possible "modes"), it is sometimes difficult for the human operator to track the activities of their automated partners. The result can be situations where the operator is surprised by the behavior of the automation asking questions like, what is it doing now, why did it do that, or what is it going to do next [Wiener, 1989]. These "automation surprises" are particularly well documented (e.g., [Degani and Heymann, 2000]; [Palmer, 1995] [Sarter and Woods, 1994][Sarter and Woods, 1995]; [Van Charante et al., 1992] and have been listed as one of the major cause of incidents (see for example [Abbott et al., 1996]. The nature of AI algorithms is likely to exacerbate the propensity of this phenomenon Rouse and Morris' definition of MM collides with a specific property of novel AI techniques known as the transparency/performance trade-off [Gunning and Aha, 2019, Figure 1]. Indeed, the more AI models are able to seize the nuance and abstraction of the data they aim to value, the less their internal states and outputs are understandable by humans. The observation of this incompatibility between AI models and their intended use as AAs has led to the emergence of eXplainable Artificial Intelligence subfield within AI research. Though, XAI approach today suffers of a lack of openness to other sciences which should be necessarily involved for answering fundamental questions such as "what has to be explained and how ?". This lack of multidisciplinary approach may be imputed to the time pressure in product release, the need for finding solutions to conform with new regulations introducing "right to explanations" [High-Level Expert Group on AI, 2019] for automatic decision systems but also by a lack of understandings of the nature of AI from some human factor specialists. This observation has been brought to the scientific community by Tim Miller in [Miller et al., 2017] who writes "this paper argues most of us as AI researchers are building explanatory agents for ourselves, rather than for the intended users." but also in the US Defense Advanced Research Project Agency (DARPA) which launched in August 2016 a new XAI project. The project scope [DARPA, 2016, II-1-B] mentions:

> "The target of XAI is an end user who depends on decisions, recommendations, or actions produced by an AI system, and therefore needs to understand the rationale for the system's decisions."

Both Tim Miller and the authors from DARPA's program have given a big contributions for answering the questions of "what should be explained and how ?".

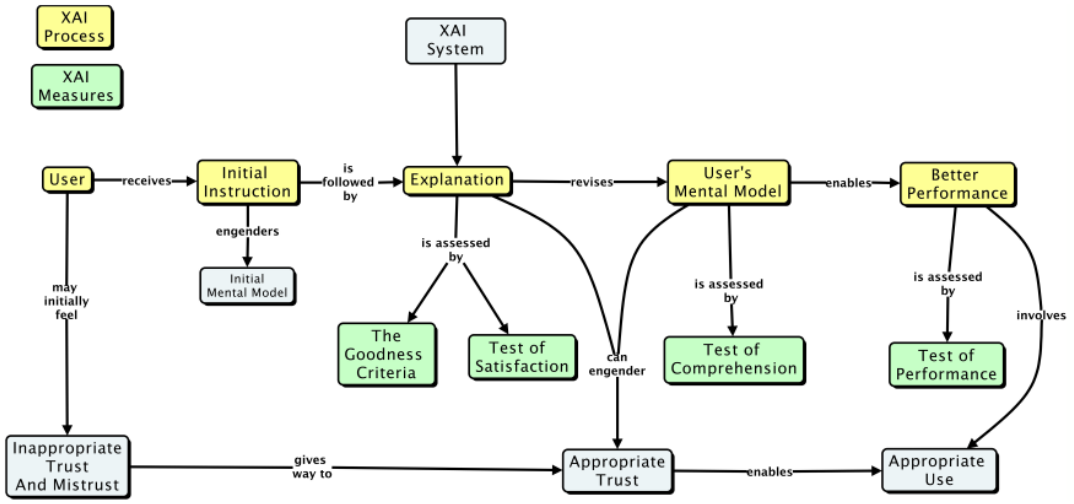

Figure 1. Conceptual model of explaining process [Hoffman et al., 2019]

2.1. Explanations as a Mental Model modulator

In [Miller, 2018] it is claimed that explanations are especially needed when the system's operator switches from inductive to abductive reasoning. According to [Aliseda, 2006], abduction (or Inference to the Best Explanation, IBE) is triggered by novelty and/or anomaly with regards to the prevailing theory. This is consistent with the conceptual model of explaining process proposed by [Hoffman et al., 2019] (see Figure 1) where explanations' role is to revise the system operator mental model of the system ("user's understanding of the AI system", [Hoffman et al., 2019]). In the case of a system operator, abduction triggers can be extended to a behavior inconsistent with the representation the operator has of the system he is interacting with. In such cases, the operator will browse hypotheses for an explanation. This is, according to Hoffman's model and Miller argumentation, where the explanation system should intervene.

2.2. Designing explanations to maintain an accurate mental model

As the need for explanation is indirectly triggered by a gap between the facts and the mental model of the user, [Miller, 2018] therefore argues that explanations should always be *contrastive*. Indeed, [Hesslow, 1988] and [Lipton, 1990] underline that when going up the causal tree of an event, relevant causes for explanations are the exceptions and should be selected regarding their pivotal role between what would have happened with the prevailing theory (*foil*) and what actually happened (*fact*). Therefore, an efficient explanation with respect to the contrastive explanation model answers the question "why the fact is observed instead of the foil ?". Such contrastive explanation should revise the user's mental model of the system, and according to Hoffman give way to an appropriate trust and therefore an appropriate use of the system assessed by better performance. [Hoffman et al., 2019] then proposes an explanation satisfaction scale and methods for eliciting mental models. [Miller, 2019] reviews publications about bias in explanation selection. Moreover, [Miller et al., 2017] claims that it is crucial to integrate the subjectivity when dealing with explanations. Indeed, [Lombrozo, 2007] shows in an experimental study how participants tend to select simpler causes over mathe-

matically more likely ones (complexity bias). Also, [McClure et al., 2007] found that in an activation chains "A allows B implying C" where participants where asked for the cause of C, participants significantly preferred intentional over natural causes , no matter if they were the activation event A or the direct event B (intentionality bias):

> "A person fanning flames is seen as a better explanation of a forest fire than the wind fanning the flames, even though the forces at work are equivalent in terms of their effects on the probability of the outcome" [McClure et al., 2007]

Finally, [Kahneman et al., 1982] and [Hilton and Slugoski, 1986] showed that participants prefer abnormal causes over more frequent causes in situations where objectively both had the same influence (abnormality bias). These results raise a new challenge while designing explanations: for AAs, it won't be enough to select the most probable foil when designing an explanation, they will have to integrate the specifics biases of the individuals they are dealing with. Another difficulty in providing relevant explanations is shown by [Mueller and Klein, 2011]: the need of explanations depends on the operator experience with the system.

3. From decision aid systems to cooperating artificial agents

The previous section about XAI has brought precious knowledge about why explanations are needed and how they should be given to an AI system operator. But this knowledge is really centered on decision-aid systems already (or about to be) on the market. In the scope of these products, the system is not considered as an agent, thus explanations are not thought as cooperation enablers. The work in the XAI paradigm is also relevant regarding the postulate that future systems including AI classification/regression modules should meet the same acceptability and trust levels as classical systems, entrusted with formal verification. This postulate is absolutely appropriate when it is about improving existing formally-verified systems with machine learning modules (for example, automatic landing systems for planes). Although, considering changing the approach while designing ML-embedding systems intending to achieve cooperation which could not be achieved with formally verified systems can be argued.

As briefly described in the introduction, AAs aim at accomplishing tasks *on behalf* humans, among them and having influence on objectives they share with them. Easy examples are autonomous cars, which will not be the norm overnight and will very likely have to share the road with human-driven cars for a time. In this context, human drivers and autonomous car will at least share the goal of avoiding accidents. This is a *cooperation* situation, with the meaning of "a process whereby two or more individuals work together toward the attainment of a mutual goal or complementary goals" (APA).

In this case of autonomous systems which considered as autonomous agents, the entity {operator, AA} could be considered - and so evaluated - as a social system {operator_1, operator_2} would be on the same cooperation task. In other words, while we consider entrusting to AA tasks which can only be achieved by humans so far, the interaction between humans and artificial stakeholders in the task must be enlightened by the interaction which would have appeared between human stakeholders, including in the performance evaluation. In this paradigm, explanations are not only supporting trust and acceptability, but are also task achievement and performance driving forces. Miller's and DARPA contributions must be taken into account while designing the format of explanations, but more knowledge about cooperation mechanisms must be

included while designing the whole interaction between human and autonomous agent cooperating.

3.1. Levels of automations : the need to share control

Researchers communities - HCI, but also Human-Machine System (HMS) and Systems, Man and Cybernetics (SMC) communities for example - have been exploring solutions for designing efficient human- systems interaction for decades. Two distinct but complementary questions have emerged (for a review see [Flemisch et al., 2019]) : 1/ how to share control between agents, 2/ how to ensure effective coordination between the different agents?

Historically, the HCI community has been interested in the issue of shared control, notably through the concept of levels of automation or LoA. Critically, automation is not a unitary concept and there is not just one way to implement automation. Automation refers to the full or partial substitution of a task or function initially performed by the human operator. As pointed by Paul Scharre, "it is meaningless to refer to a system as 'autonomous' without referring to the specific task that is being automated. For any given task, there are degrees of autonomy" [Scharre and Horowitz, 2015]. In that sense, automation is not all or none but can vary across a continuum of levels. Each of these classes of function can be automated on different levels, from "human does everything manually" to "machine does everything and ignore human". Following this rationale, different scales of levels of automation (LoAs) have been proposed (see for example [Sheridan and Verplank, 1978]; [Nuhuis, 1999]; [Kaber and Endsley, 2004]).

Starting from this premise, a classical approach was to manipulate the LoAs in order to improve human automation interaction. In particular, a very classical method consisted in sharing the tasks to be carried out between man and machine according to the strengths and weaknesses of each. MABA-MABA lists, or 'Men Are Better At–Machines Are Better At' lists have appeared over the decades in various guises (e.g. [Chapanis, 1965]; [Mertes and Jenney, 1974]; [Swain and Guttmann, 1983]; [Sheridan, 1987]). Later, similar design advice was offered, but with a focus on functions rather than tasks. For example, [Parasuraman et al., 2000] propose that designers divide the tasks between humans and machines by considering four different groups of system functions: Information acquisition, Information analysis, Decision and action selection, Action implementation. A simple flowchart is presented that leads the engineer from the question "What should be automated?" to identifying the types of automation (one choice from the four functions above). Then the engineer can choose from a list of automation levels. Then, in classic MABA-MABA style, the advantages and disadvantages of automating parts of each of the four functions are discussed. Similarly, other authors have proposed to decompose tasks into operational, tactical and strategic levels (e.g., [Abbink, 2006]; [Lemoine et al., 1996]; [Woods et al., 2004]). Interestingly, these different taxonomies (see also the taxonomy in [Kaber and Endsley, 2004]) provide a systematic way to discover the important aspects of automating a task, based on a clear link to cognitive theory.

Although these taxonomies are slightly different - because they are based on a detailed breakout of the aspects of human performance that are being automated - it has been possible to conduct research to determine much more precisely how the addition of automation to various functions affects human performance and to address the difficulties encountered by human operator interacting with automated systems. A central

finding is that intermediary levels of automation could maintain operator involvement in system performance, leading to improvements in situation awareness and reductions in out-of-the-loop performance problems ([Endsley et al., 1997]; [Manzey et al., 2012]; [Metzger and Parasuraman, 2001]; [Endsley, 2018b]). In summarizing the LoA literature, [Endsley, 2017] developed a model of LoA effects on operator engagement and workload that can help to illuminate the design trade-offs involved. However, these techniques have also received a lot of criticism, in particular because it does not take into account the way in which the introduction of automation at different levels can alter, modify the very nature of the work carried out by the human operator (see [Dekker and Woods, 2002]). In this method, system automation is considered as a simple substitution of a machine activity for human activity, a belief called "substitution myth" [Woods and Tinapple, 1999]. Unfortunately, such assumption corresponds to a distorted reflection of the real impact of automation: automation technology transforms human work and forces people to adapt their skills and routines [Dekker and Woods, 2002]. Whatever the merits of any particular automation technology are, automation does not merely supplant human activity but also transforms the nature of human work.

Particularly, creating partially autonomous machine agents is, in part, like adding a new team member. One result is the introduction of new coordination demands and the emergence of new classes of problems which are due to failures in the human-machine relationship. Many of the challenges faced by designers of human-machine interactions involve teamwork rather than the separation of tasks between human and machine [Klein et al., 2004]. Effective teamwork involves more than an efficient division of labor; it seeks ways to support and enhance each member's performance - a need not addressed by the levels of autonomy conceptualization, nor by adaptive automation methods.

3.2. Beyond sharing control on actions: escalating layers towards cooperation

This issue of cooperation amongst team (and/with automates) has led research on (team) performance and situation awareness investigating how system might support collaboration between operators. Initially, [Dekker and Woods, 2002] proposed several principles to shape how information about automation and the processes in controls are displayed to the operator to enhance human-automation teaming: highlighting changes, displaying future projections, and visually integrating information (see also the "ten challenges for making automation a team player" proposed by Klein and collaborators [Klein et al., 2004].

Since then, the issue of cooperation has been widely investigated and many works underline the importance of taking it into account when designing artificial agents ([Banks et al., 2013]; [Flemisch et al., 2016]; [Hoc, 2001]; [Hoc et al., 2009]; [Millot and Lemoine, 1998]; [Hutchins, 1995]).These frameworks highlight the importance of the information processing and the communication between human operators and automated systems to create a shared representation. For instance, it has been proposed to design automation systems as chatty co-drivers providing continuous relevant feedback to the driver to improve human automation interaction ([Eriksson and Stanton, 2015]; [Stanton et al., 2011]).

Regarding human-machine cooperation, [Hoc and Lemoine, 1998] observes that cooperating partners "interfere" positively or negatively at goal's place like physical

waves and extends APA's definition of cooperation by including the will for a cooperator "to make the activities of the other (partners) easier

Another interesting framework has been recently proposed by Pacaux and collaborators [Pacaux et al., 2011]. These authors distinguishes between the "know-how" (to operate) and the "know-how-to-cooperate" both essentials to the cooperating agents. Agent's know-how refers to its task-related skills at diagnosis, decision or action levels while agent's know-how-to-cooperate covers its ability to identify what its partners are doing or planning by "interference" management at action level. This know-how-to-cooperate extends itself at plan level by elaborating or maintaining a common frame of reference regarding "common goals, common plans, role allocation, action monitoring and evaluation, and common representations of the environment" and at a meta-model -allowing meta-cooperation - providing agents with models of the other agents. Critically, the ability to cooperate relies in this framework on the existence of a "common work space" which provides the operator with information about his own environment and action but also about the current and future actions of other agents (see also [Pacaux-Lemoine and Debernard, 2000]). This common work space aims to provide a shared mental model, here called team Situation Awareness [Millot and Pacaux-Lemoine, 2013]. This concept of shared mental model has been largely used in the context of team work ([Converse et al., 1993]; [Mathieu et al., 2000]). The idea is that mental models are necessary for team members to predict what their teammates will do and what they will need, thus facilitating the coordination of actions between teammates. In this way, mental models help to explain team functioning. This line of research has provided interesting concepts and methods: team SA [Salas et al., 1995] [Gorman et al., 2006], distributed cognition framework [Hutchins, 1995]; [Stanton, 2016] or adaptive automation [Miller and Parasuraman, 2007]. Interestingly, several studies have shown a positive relation between team performance and similarity between mental models of team members (see, e.g., [Bolstad and Endsley, 1999];[Mathieu et al., 2000]; [Lim and Klein, 2006]. The next subsection will show how these concepts and principles are being completed in another community working specifically on human autonomy teaming.

3.3. Human Autonomy Teaming

Recently, another scientific community driven by AI progress has been working on Human-Autonomy Teaming (HAT). According to [McNeese et al., 2018], the progresses in AI tends to make autonomous systems more *intelligent*. For this reason they should no longer only be considered as *servants* but as *teammates*. Along those lines, McNeese considers Human-Machine Interaction (HMI) has to be extended by a new scientific field he calls Human-Automation Teaming (HAT). In 2012, [Johnson et al., 2012] already suggested to switch from the paradigm of AI systems accomplishing tasks *for* humans to AAs working *with* humans. The authors of this paper showed in a study (using opensource Blocks World for Teams (BW4T) testbed) that increasing AAs' performance and autonomy is not enough for reaching better performances on collaborative tasks with humans. [McNeese et al., 2018] completes these results in the CERTT UAS-STE testbed [Cooke and Shope, 2002]: though the overall performance is similar between groups, human-autonomy groups presents less efficient targeting than the control group (human-human) and analyzing the communications shows less information push and more pulls from the human teammate perspec-

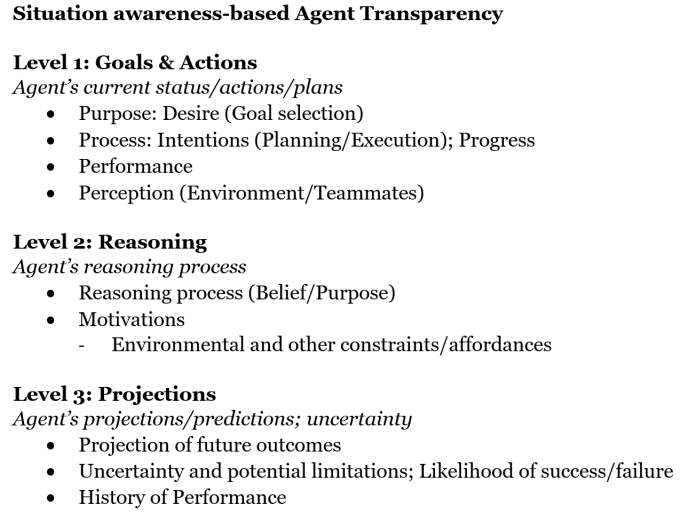

Figure 2. Situation Awareness-based Agent Transparency model [Chen, 2018, Figure 1]

tive. Based on these results, HAT papers [O'Neill et al., 2020] [McNeese et al., 2018] [Chen et al., 2014] [Grimm et al., 2018] have been pushing two key variables as co-operation enablers (or mediators, acknowledging the Input-Mediator-Output model from [Kazi et al., 2019, Figure 1]): shared mental models and team situation awareness (TSA). [Gorman et al., 2006] is the reference for the HAT community regarding TSA (also introduced in [Salas et al., 1995]). This paper itself refers to [Endsley and Jones, 1997, p17] for situation awareness definition:

> "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future"

but reject the "comprehension of their meaning" part for breaking the continuous "perception-action" process they mean by SA. Then SA is extended to TSA in the same paper (p46):

> "the degree to which every team member possesses the SA required for his/her job"

Hence, [Chen et al., 2014] proposes Situation Awareness-based Agent Transparency (SAT) model. Figure 2 reproduces the updated SAT model from [Chen, 2018]. Empirical studies using this model brought meaningful results: [Mercado et al., 2016] shows that increased transparency leads to increased performance on task "without additional costs" (on effectiveness and time). [Lyons, 2013] also studies the needed transparency in HRI for human SA and developping a "shared awareness" with robots. The paper highlights behavioral models which should be transparent to humans : intentional model, task model, analytical model, environment model and teamwork model. This is an interesting precision of the notion of mental model presented earlier (see section 2.1). Some of the suggested information such as the information on the logic behind system's decision [Lyons et al., 2017] - referring to the analytical model - has been proven to improve trust in user-studies [Lyons et al., 2017]. As an extension of the wide spread "mental model" concept, the American Psychological Association defines Shared Mental Models (SMM) as

> "in ergonomics, a mental model of a work system that is held in common by the members of a team. Ideally, team members should have a shared mental picture of the system and

its attributes, a shared knowledge of all relevant tasks, and a shared understanding of the team's progress toward its goal. Coordination, efficiency, and accuracy will increase as team members converge on a common mental model that is accurate and complete yet flexible. Also called team mental model."

SMM and TSA are two key intricated concepts of the teaming dynamic in ergonomic [Ososky et al., 2012], they are powerful high-level analysis tools for human-AA cooperation. Because these tools have been used for at least 3 decades, measurements have been tried and tested and concrete impact within HMI demonstrated [Endsley, 2018a]. Although, some authors point out the need of giving interest to lower-level psychological mechanisms in order to improve SMM and TSA:

"The most serious shortcoming of the situation awareness construct as we have thought about it to date, however, is that it's too neat, too holistic and too seductive. We heard here that deficient SA was a causal factor in many airline accidents associated with human error. We must avoid this trap: deficient situation awareness doesn't cause anything. Faulty spatial perception, diverted attention, inability to acquire data in the time available, deficient decision-making, perhaps,but not a deficient abstraction!"[Billings, 1995]

"This review points out confusion surrounding the concept and use of mental models from the viewpoints of both human factors and psychology."[Wilson and Rutherford, 1989]

At the end of this section, a first assessment can be drawn: first, the literature relevant to the problem of task sharing and spatial coordination between humans and AAs is rich, the problem has been stated and addressed from multiple angles for at least three decades. Then this literature is coherent, though lacking of of semantic unity. For example the notions of meta-model - belonging to the know-how-to-cooperate - from [Pacaux et al., 2011] must be read in parallel with the HAT research as they may encompass several concepts and ideas. Once this has been said, know-how-to-cooperate, interference management, SMM and TSA can be considered as "desirable states" or meta-mechanisms which AAs design must must look for. Finally, the notions presented in this section are also coherent with the research presented in section 2 while bringing a lot of precision and details: mental models keep holding a key role while interferences echo the notion of contrastive explanations presented in subsection 2.2. The next section will describe lower-level cognitive mechanisms "of interest" for observing TSA and SMM during AA-human cooperation.

4. Cognitive mechanisms for shared-mental models and team situation awareness

Previous sections allow to specify the requirements to reach in cooperative-AA design in scientific terms: AAs must be at least able to provide their human partners with an appropriate mental model of their behavior (to generate a SMM), a team situation awareness and being able to provide contrastive explanations including human biases. One can argue that the Chen's SAT model and Lyons' model give precise lists of the required abilities for an AA cooperating with humans as well as the information needed to acquire these abilities. However, it is reasonable to think that the performance-transparency trade-off of AI techniques will compromise the availability of some required information described in the model. Moreover, as human-human

cooperative interactions do not require such formalism, we argue that the study of *necessary and sufficient* information activating finer level cognitive mechanisms involved in cooperation is of particular interest. In cognitive science, the study of cooperation can be found in joint-action and intentional stance literature. If the concepts of mental models and situation awareness are not mentioned, following descriptions will show that both approaches are complementary.

4.1. Shared Task Representation

[Knoblich et al., 2011] introduces joint-action by "When two or more people coordinate their actions in space and time to produce a joint outcome, they perform a joint action." and then distinguishes between emergent and planned coordination.

> "In planned coordination, agents' behavior is driven by representations that specify the desired outcomes of joint action and the agent's own part in achieving these outcomes. [...] In emergent coordination, coordinated behavior occurs due to perception–action couplings that make multiple individuals act in similar ways; it is independent of any joint plans or common knowledge (which may be altogether absent)."[Knoblich et al., 2011]

To our understanding and given the definition of cooperation we adopt, planned coordination is necessary in human/AA cooperation. Knoblich and his co-authors have found that in effective planned coordination situations, agents performing joint-action represents both their own task and their co-actors': agents have a shared task representation. Shared task representation between human co-actors is not only supported by behavioral studies (social Simon effect, [Sebanz et al., 2005]), but also with brain imaging and electrophysiological studies [Ramnani and Miall, 2004] [Sebanz et al., 2006]. In [Sebanz et al., 2005], the authors perform experiments in order to show that participants involved in joint action do not only represent their co-actor's actions, but also their co-actor's task. The demonstration starts from Simon's finding about slower response times (RTs) from participants to a color stimulus when an irrelevant spatial information is presented whether than a relevant, e.g.: participants have a green button under their left hand they must press when a green stimulus is showed on a screen, and a red one under their right hand they must press when a red stimulus is showed on the screen. Simon's results [Craft and Simon, 1970], [Simon, 1990] show that there is a significantly slower RTs when the green (respectively red) stimulus is presented incompatibly on the right (respectively left) side of the screen than on the compatible left side (respectively right). In [Sebanz et al., 2003], Sebanz and colleagues showed that this effect is also social by measuring significantly slower RTs with both compatible, neutral and incompatible spatial information when a go/nogo action (e.g. in response to a green stimuli) is performed spatially along someone performing the same action in response to another color (e.g. red) than when the same go/nogo task is performed alone. This suggests that when people think being engaged in joint-action, they at least share a representation of their actions. Finally, in [Sebanz et al., 2005], the authors want to elicit whether participants engaged in joint-action simply share actions representations or also task representations: in this experiments, a participant sitting left must respond to a spatial stimulus (finger pointing left) while another one sitting right must respond to a color stimulus (green ring on the finger). The experiment shows significantly slower RTs in "double response situation" - meaning that both the finger was pointing left and the ring green - than when only one of the two participants had to respond (either only finger pointing left with red ring or green ring with finger pointing right). This result suggests that co-actors do not only

represent their partner's action, but also their partner's task. It was later enforced with electroencephalography (EEG) studies : [Kourtis et al., 2013] shows by using an electrophysiological evidence (contingent negative variation, CNV) that "joint action planning involves cognitive and motor representations of the action partner's task" and later in [Kourtis et al., 2014] shows that CNV is similar when asking participants to clink two glasses by themselves or to clink one glass with another participant's glass.

Although - and relevantly to our problem - experiments in [Tsai et al., 2008] confirmed in [Sahaï et al., 2017] shows that shared task representations evidence are only observed when participants are told that their co-actor is human (as opposed to "algorithm"). Indeed in their experiment Tsai and colleagues proposed a joint Simon task to the participants. Participants where asked to respond to green stimuli with compatible, neutral or incompatible - regarding participant's seat position - position on the screen. Then they were told that either a friend (condition 1) or an algorithm (condition 2) would take care of the response to the red stimuli, while actually all the red stimuli were managed by the computer. The study shows that social Simon effect (significantly slower responses for incompatible stimuli) was only observed in condition 1, meaning that shared task representation mechanism is not triggered when we think being co-acting with a computer. Meanwhile, there are electphysiological evidences for participants mentally performing their co-actor's task in condition 1. Following [Tsai et al., 2008], the cognitive ambitions for autonomous agents must include the ability of inducting shared task representation to its co-actors.

4.2. Mentalization

Mentalization is defined as "the ability to understand one's own and others' mental states, thereby comprehending one's own and others' intentions and affects."(APA). It is considered as *metacognition applied to others* where metagonition is "the processes by which we monitor and control our own cognitive processes"[Frith, 2012]. Mentalization is considered to participate into intentional stance ("a strategy for interpreting and predicting behavior that views organisms as rational beings acting in a reasonable manner according to their beliefs and desires (i.e., their intentions)"[Dennett, 1989]) which appears to be a requirement to social interactions, including cooperation [Dennett, 1971].

In the seek for shared mental models between AAs and human partners, mentalization appears by nature to be a cognitive mechanism of interest. Wako Yoshida while studying Theory of Mind (ToM) - with the narrow definition of "how we represent the intentions and goals of others to optimise our mutual interactions", definition equivalent to mentalization - provided precious results about mentalization towards AAs. In a first experiment in 2008 [Yoshida et al., 2008], two hypothesis about participants behavior were tested in a Sequential Stag-Hunt (SSH) game: ToM or not ToM, meaning whether the participants infer the behavior of their co-actor in their decision or not. SSH is derived from the game-theory situation of the stag-hunt, where participants must chose between a cooperative but risky strategy - as if the co-actor doesn't chose to cooperate, everything is lost - or a selfish but sure strategy, guaranteying a minimum reward no matter the choice of the co-actor. SSH is played in a maze grid and co-actors must chose between hunting cooperatively a randomly-moving stag by acting in order to find themselves at the same time on an adjacent to the stag grid-spot, or hunting selfishly a static rabbit. Co-actors groups where formed with a human participant and an artificial agent alternating between cooperative and self-

ish policies. Participants behavior were then compared to behavior they would have adopted with a fixed policy or a policy inferring the artificial agent intentions from his behavior. Results showed evidence that human co-actors adapt their representations of their co-actors depending of their actions. In a second experiment, a behavioral study shows that the sophistication of our reasoning, meaning how many step forward are integrated in the decision making during the stag-hunt is correlated to the co-actor's sophistication itself: the more our co-actors takes our intentions into account, the more we are willing to infer theirs [Yoshida et al., 2010]. Furthermore, the study is completed with brain imaging showing that the brain parts associated in the literature with ToM are actually activated by the subjects during the task. Although according to [Perez-Osorio and Wykowska, 2020], the adoption of intentional stance towards AAs is not natural. The paper mentions studies including Prisoner's Dilemma game and Rock-Paper-Scissor games with opposite results to Yoshida and colleagues' regarding brain imagery. The authors argue that human-like appearance of the agents, the context, goal-oriented behavior as well as individual priors affect the adoption of intentional stance towards AAs.

5. Hypotheses of informations to share in order to trigger shared-task representation and mentalization

5.1. Intention sharing as shared-task representation and mentalization enabler

The two bodies of literature of joint-action - bringing shared task representation - and intentional stance have brought cognitive mechanisms to the seek of shared mental models and team situation awareness. Although, both mechanisms have been shown to be hard to activate towards computers or robots. To our understanding it is a crucial requirement for AAs to be able to arise these mechanisms on their human partners in action. Because AAs are not necessarily embodied robots, the content and the format of the interaction between AAs and humans in cooperation - directed towards STR and intentional stance - has to be found in the underlying mechanic of the AAs. The literature leads us to the fundamental notion of intentions. The DPM (Distal (or future-directed)-Proximal (or present-directed)-Motor) model of intentions [Pacherie, 2008] first has to be introduced:

> "F-intentions are formed before the action and represent the whole action as a unit. They are usually detached from the situation of action and specify types of actions rather than tokens. Their content is therefore conceptual and descriptive. [...] P-intentions serve to implement action plans inherited from F-intentions. They anchor the action plan both in time and in the situation of action and thus effect a transformation of the descriptive contents of the action plan into perceptual-actional contents constrained by the present spatial as well as non-spatial characteristics of the agent, the target of the action, and the surrounding context. The final stage in action-specification involves the transformation of the perceptual-actional contents of P-intentions into sensorimotor representations (M-intentions) through a precise specification of the spatial and temporal characteristics of the constituent elements of the selected motor program." [Pacherie and Nicod, 2007]

At this point - and keeping in mind the importance of demonstrating the coherence in the available literature - it is interesting to note the compatibility between distal intentions and Lyons' intentional model as well as between proximal intentions and task model [Lyons, 2013].

Abilities	**Mechanisms**
1 - Shared perceptual representations of the situation of action	- joint attention (causal coordination & mutual manifestness)
2 - Corepresenting the actions and proximal intentions of other agents	- perception-action interface: motor resonance via mirror neurons supporting action understanding and outcome prediction (goal-to-action) - action anticipation via goal-to-action + inference from teleological reasoning -task sharing
3 - Having an accurate representation of the system's dynamic to allow triadic adjustment	"agents should be capable of explicitly representing the instrumental relation of their individual actions to the situated joint goal structure"

Table 1. 3 types of abilities the co-agents must be capable of and the proposed mechanisms after [Pacherie, 2012]

In [Pacherie, 2012], Elisabeth Pacherie extends the DPM model to joint-action. The existence of Shared Distal Intentions (SDI) derives from the intrinsic nature of planned joint action which implies sharing the overall goal and including cooperators influence on one's subplan sketches. The paper then connects recent empirical evidences to Shared Proximal Intentions (SPI) and Coupled Motor Intentions (CMI). There are 6 conditions characterizing SPI:

- self-prediction
- other-prediction
- dyadic adjustment (representation on one's action on others')
- joint action plan
- joint predictions
- triadic adjustment (joint progress monitoring and next move decision, including moves which could help partner to achieve their contribution)

To observe these characteristics in an interaction, Pacherie relying on [Sebanz et al., 2006] lists three types of abilities the co-agents must be capable of and the proposed mechanisms. They are resumed in Table 1. Regarding Coupled Motor Intentions, Pacherie considers their importance is correlated to the precision of the space-time coordination needed in the task. Mentioned example include dancing ballets and rowing. In this case, the entrainment mechanism proposed in emergent coordination by [Knoblich et al., 2011] is proposed as a key to CMI, also explaining why the author does not talk about *shared* but *coupled* motor intentions as it is more dyadic than triadic adjustment.

We suggest that shared intentions must be further explored regarding their potential to be facilitators towards shared task representation and mentalization in human-AA cooperation. AS illustration, Le Goff and colleagues (2018) recently showed that communicating information about the system's intention to act improved the operator sense of control and performance as well as system acceptability.

5.2. Metacognitive representations as mentalization triggers

The studies mentioned in section 4.2 at least suggest that mentalization towards AAs is neither natural nor automatic. In order to model the mental states of a partner (mentalizing) after adopting toward him the intentional stance - meaning considering the partner as an intentional agent - this partner has to provide clues about its own mental processes. Although, the nature of these clues remain unclear [Johnson, 2003] [Mar and Macrae, 2008].

One of this clue could be - as mentalization is *metacognition applied to others* [Frith, 2012] - the ability to share metacognitive information. [Shea et al., 2014] argues that humans are provided with a "system 2 metacognition" dedicated to suprapersonal cognitive control which allows the sharing of metacognitive information. This system 2 metacognition would thereby permit "to coordinate the sensorimotor systems of two or more agents involved in a shared task" by transmitting explicit metacognitive representations about ongoing action process. Moreover communicating about metacognitive representations is known to make the joint-action smoother [Lausic et al., 2009] and to contribute to joint performance [Bahrami et al., 2012]. Among the possible metacognitive representations, the confidence regarding its proposed outputs is of particular interest by being technologically plausible regarding the current trend in AI. Using an avoidance task, we have recently shown that the metacognitive information provided by the system (i.e., its own confidence in the solution proposed) helped restore the operator's sense of control, and this increased in feeling of control was also associated with greater system acceptability [Vantrepotte et al., 2022].

6. Conclusion

The two first sections of this paper commits into building a consistent picture of the disparate concepts from ergonomic, cognitive science and XAI relevant to the question of interaction requirements for AAs aiming at cooperating with humans. Understanding that most of these concepts are complementary rather than incompatible is reassuring: cooperating humans and AAs should share their mental models and hold the same perception and interpretation of the situation. These findings question about the possible cognitive mechanisms contributing to these desirable ergonomic states: While the trends in AI drive the research to always less constrained models, lighter design requirements coming from interactions abilities will be the more likely to be applied in practice. The literature on joint action involves at least two candidate mechanisms: shared-task representation and mentalization which are both demonstrated as unnaturally triggered in joint-action between humans and AAs, raising a new challenge.

This challenge is handled in the last section of the paper by proposing credible hypotheses of content to be shared by artificial agents engaged in joint-action with human partners in order to activate shared task representation and mentalization. These hypotheses focus on intention sharing using [Pacherie, 2012]'s model of intention is a joint-action context and the sharing of metacognitive representations (including confidence level on the AA's outputs). The reasoning process of this paper is summed-up in the figure 3. Next step will include to measure the influence of intentions sharing in cooperation. This implies to integrate the influence of each DPM level of intentions, to evaluate to what extent the shared task representation mechanism is involved, and to measure team situation awareness and shared mental models. This work must also

as much as possible be enlightened by the knowledge from XAI about contrastive explanations and bias.

Though the idea of manned and unmanned aerial or terrestrial vehicles working jointly on search and rescue or firefighting situations being the starting point of our research, this article is intended to be as generic as possible regarding human - AA cooperation. This way, the article may provide AI engineers working on AA models with the key principles of cooperation and assumptions about how the interaction can be smoothed and improved. However, many questions remain to be addressed. For example, does the proposed approach depend on whether the human-system relationship is symmetric or not? Most of the work on joint actions in humans refers to symmetrical situations, whereas the relationship between the human and the AI is asymmetrical in most of the cases. Our recent work ([Le Goff et al., 2018]; [Vantrepotte et al., 2022]) seems to indicate that the proposed approach is not dependent on the symmetry of the human-system relationship. Particularly, our first studies confirm the relevance of this approach for two distinct situations in terms of the relative involvement of the human operator and the automatism in the produced action. However, in both cases, the information transmitted by the system - whether it is information related to its intentions [Le Goff et al., 2018] or to the confidence in the proposed solutions [Vantrepotte et al., 2022] - generate the same beneficial effect in term of experience of control and system acceptability. Nevertheless, one might wonder whether this information would continue to have a positive impact in situations of extreme complexity where the human operator could hardly complement the actions of the AI.

Symmetry of the interaction is a complementary problem : to remain realistic, the interaction hypotheses presented in the paper should stay in line with AI techniques. As for now the interpretation of human joint action signals (natural language, body language, gaze...) remains a challenge for computers, this paper mainly focuses on signals coming from AAs towards humans. Though communicating intentions adds constraints to the agent's design, this remains realistic regarding AI state-of-the-art (see [Vezhnevets et al., 2017] which hierarchichal reinforcement learning approach matches some intentions communication requirements). Mechanically, being able to extract such information from the agent's model affects its "black-boxness" and - referring to the transparency/performance trade-off - may affects its performance along with bringing constraints to engineers and researchers.

A second question refers to the applicability of these interaction principles to the different situations that may be encountered by operators. It is clear that we are dealing here with cooperative situations, in particular when it comes to producing coordinated behaviour or making so-called collaborative decisions. In these conditions where the mutual understanding of each partner's actions is relevant to the task, it is strongly expected that the proposed principles will improve the performance of the human-AI system. However, there are many other situations where the relevance of this information is not guaranteed. For example, what about emergency situations for which a negotiation in terms of decision making will not necessarily be expected. One could also question the relevance of this information sharing in situations where the human operator has a simple supervisory role (but see the results of [Le Goff et al., 2018]).

Although promising, the proposed framework requires further study to address the concerns they raise and to test it in more complex and ecological situations than the laboratory studies traditionally used.

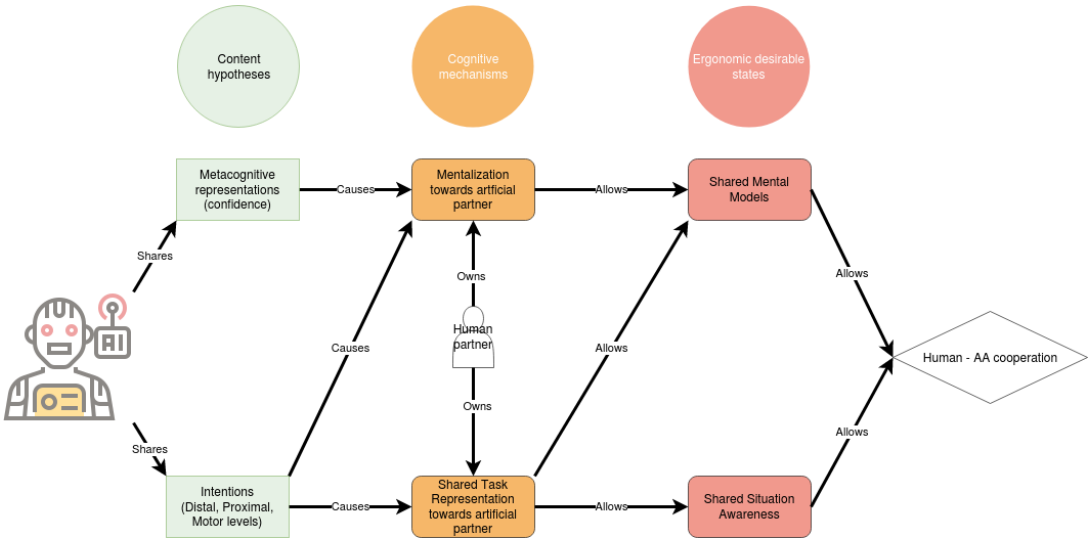

Figure 3. A summary of the reasoning held in the paper

Disclosure statement

We have no relevant financial or nonfinancial relationships to disclose.

Funding

This research project is supported by ONERA the French aerospace lab and Region SUD

Biographies

Marin Le Guillou

Marin Le Guillou is a PhD student at ONERA and LPL working on Human-Artificial Agents cooperation towards a smooth integration of manned and unmanned drone engaged in cooperative tasks. He graduated in artificial intelligence of Centrale Marseille engineering school and Aix-Marseille University.

Bruno Berberian

Dr. Bruno Berberian is a senior researcher with 10 years of experience in cognitive ergonomics. He is currently working as a research engineer at the ONERA. His current research investigates how automation technology impact human operator and aim to propose tools and models from cognitive science to explore this issue.

Laurent Prévot

Laurent Prévot is a professor in Language Sciences at Aix Marseille Université. He obtained his PhD in Computer Science and worked as post-doc in interdisciplinary

labs in Europe and Asia. His recent projects are dealing with conversational feedback and interpersonal dynamics in conversation.

References

[Abbink, 2006] Abbink, D. (2006). *Neuromuscular analysis of haptic gas pedal feedback during car following*. Dissertation, TU Delft, Delft. ISBN: 9789085592532.

[Abbott et al., 1996] Abbott, K., Slotte, S. M., and Stimson, D. K. (1996). The interfaces between flightcrews and modern flight deck systems. Technical report, United States Federal Aviation Administration. Publisher: United States. Federal Aviation Administration.

[Aliseda, 2006] Aliseda, A. (2006). *Abductive reasoning: logical investigations into discovery and explanation*. Number 330 in Synthese library. Springer, Dordrecht, The Netherlands. OCLC: ocm70052496.

[Bahrami et al., 2012] Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., and Frith, C. (2012). Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, 38(1):3–8.

[Banks et al., 2013] Banks, V. A., Stanton, N. A., and Harvey, C. (2013). What the crash dummies don't tell you: The interaction between driver and automation in emergency situations. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 2280–2285. IEEE.

[Billings, 1995] Billings, C. E. (1995). Situation awareness measurement and analysis: A commentary. In *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness*, volume 1. Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

[Bolstad and Endsley, 1999] Bolstad, C. A. and Endsley, M. R. (1999). Shared mental models and shared displays: An empirical evaluation of team performance. In *proceedings of the human factors and ergonomics society annual meeting*, volume 43, pages 213–217. SAGE Publications Sage CA: Los Angeles, CA. Issue: 3.

[Carberry, 1988] Carberry, S. (1988). Modeling the user's plans and goals. *Computational Linguistics*, 14(3):23–37.

[Carroll et al., 2020] Carroll, M., Shah, R., Ho, M. K., Griffiths, T. L., Seshia, S. A., Abbeel, P., and Dragan, A. (2020). On the Utility of Learning about Humans for Human-AI Coordination. *arXiv:1910.05789 [cs, stat]*. arXiv: 1910.05789.

[Chapanis, 1965] Chapanis, A. (1965). On the allocation of functions between men and machines. *Occupational Psychology*, 39(1):1–11.

[Chen et al., 2014] Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., and Barnes, M. (2014). Situation Awareness-Based Agent Transparency:. Technical report, Defense Technical Information Center, Fort Belvoir, VA.

[Chen, 2018] Chen, J. Y. C. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, page 25.

[Choudhury et al., 2019] Choudhury, R., Swamy, G., Hadfield-Menell, D., and Dragan, A. D. (2019). On the Utility of Model Learning in HRI. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 317–325, Daegu, Korea (South). IEEE.

[Converse et al., 1993] Converse, S., Cannon-Bowers, J. A., and Salas, E. (1993). Shared mental models in expert team decision making. *Individual and group decision*

making: Current issues, 221:221–46. Publisher: Hillsdale.

[Cooke and Shope, 2002] Cooke, N. J. and Shope, S. M. (2002). The CERTT-UAV task: A synthetic task environment to facilitate team research. *SIMULATION SERIES*, 34(3):25–30.

[Craft and Simon, 1970] Craft, J. L. and Simon, J. R. (1970). Processing symbolic information from a visual display: Interference from an irrelevant directional cue. *Journal of Experimental Psychology*, 83(3, Pt.1):415–420.

[DARPA, 2016] DARPA (2016). Broad Agency Announcement Explainable Artificial Intelligence (XAI).

[Degani and Heymann, 2000] Degani, A. and Heymann, M. (2000). Pilot-autopilot interaction: A formal perspective. *Abbott et al.[1]*, pages 157–168. Publisher: Citeseer.

[Dekker and Woods, 2002] Dekker, S. W. A. and Woods, D. D. (2002). MABA-MABA or Abracadabra? Progress on Human-Automation Co-ordination. *Cognition, Technology & Work*, 4(4):240–244.

[Dennett, 1971] Dennett, D. C. (1971). Intentional Systems. *Journal of Philosophy*, 68(4):87–106.

[Dennett, 1989] Dennett, D. C. (1989). *The intentional stance*. MIT press.

[Endsley and Jones, 1997] Endsley, M. and Jones, W. M. (1997). Situation Awareness Information Dominance & Information Warfare. Technical report, LOGICON TECHNICAL SERVICES INC DAYTON OH.

[Endsley, 2017] Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned From Human–Automation Research. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 59(1):5–27.

[Endsley, 2018a] Endsley, M. R. (2018a). Expertise and Situation Awareness. In Ericsson, K. A., Hoffman, R. R., Kozbelt, A., and Williams, A. M., editors, *The Cambridge Handbook of Expertise and Expert Performance*, pages 714–742. Cambridge University Press, 2 edition.

[Endsley, 2018b] Endsley, M. R. (2018b). Level of Automation Forms a Key Aspect of Autonomy Design. *Journal of Cognitive Engineering and Decision Making*, 12(1):29–34.

[Endsley et al., 1997] Endsley, M. R., Mogford, R. H., and Stein, E. (1997). Controller Situation Awareness in Free Flight. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 41(1):4–8.

[Eriksson and Stanton, 2015] Eriksson, A. and Stanton, N. A. (2015). When Communication Breaks Down or What was that? – The Importance of Communication for Successful Coordination in Complex Systems. *Procedia Manufacturing*, 3:2418–2425.

[Finin and Drager, 1986] Finin, T. and Drager, D. (1986). GUMS: a general user modeling system. In *Proceedings of the workshop on Strategic computing natural language - HLT '86*, page 224, Marina del Rey, California. Association for Computational Linguistics.

[Flemisch et al., 2016] Flemisch, F., Abbink, D., Itoh, M., Pacaux-Lemoine, M.-P., and Weßel, G. (2016). Shared control is the sharp end of cooperation: towards a common framework of joint action, shared control and human machine cooperation. *IFAC-PapersOnLine*, 49(19):72–77.

[Flemisch et al., 2019] Flemisch, F., Abbink, D. A., Itoh, M., Pacaux-Lemoine, M.-P., and Weßel, G. (2019). Joining the blunt and the pointy end of the spear: towards a common framework of joint action, human–machine cooperation, cooperative guidance and control, shared, traded and supervisory control. *Cognition, Technology &*

Work, 21(4):555–568.

[Frith, 2012] Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2213–2223.

[Gorman et al., 2006] Gorman, J. C., Cooke, N. J., and Winner, J. L. (2006). Measuring team situation awareness in decentralized command and control environments. *Ergonomics*, 49(12-13):1312–1325.

[Grimm et al., 2018] Grimm, D. A., Demir, M., Gorman, J. C., and Cooke, N. J. (2018). Team Situation Awareness in Human-Autonomy Teaming: A Systems Level Approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1):149–149.

[Gunning and Aha, 2019] Gunning, D. and Aha, D. W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine*, 40(2):44–58. Publisher: Association for the Advancement of Artificial Intelligence.

[Hesslow, 1988] Hesslow, G. (1988). The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality*, pages 11–32.

[High-Level Expert Group on AI, 2019] High-Level Expert Group on AI (2019). Ethics guidelines for trustworthy AI. Report, European Commission, Brussels.

[Hilton and Slugoski, 1986] Hilton, D. J. and Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1):75–88.

[Hoc, 2001] Hoc, J.-M. (2001). Towards a cognitive approach to human–machine cooperation in dynamic situations. *International Journal of Human-Computer Studies*, 54(4):509–540.

[Hoc and Lemoine, 1998] Hoc, J.-M. and Lemoine, M.-P. (1998). Cognitive Evaluation of Human-Human and Human-Machine Cooperation Modes in Air Traffic Control. *The International Journal of Aviation Psychology*, 8(1):1–32.

[Hoc et al., 2009] Hoc, J.-M., Young, M. S., and Blosseville, J.-M. (2009). Cooperation between drivers and automation: implications for safety. *Theoretical Issues in Ergonomics Science*, 10(2):135–160.

[Hoffman et al., 2019] Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2019). Metrics for Explainable AI: Challenges and Prospects. Technical Report arXiv:1812.04608, arXiv. arXiv:1812.04608 [cs] type: article.

[Hutchins, 1995] Hutchins, E. (1995). *Cognition in the Wild*. A Bradford Book, Cambridge, MA, USA.

[Johnson et al., 2012] Johnson, M., Bradshaw, J. M., Feltovich, P., Jonker, C., Riemsdijk, B. v., and Sierhuis, M. (2012). Autonomy and interdependence in human-agent-robot teams. *IEEE Intelligent Systems*, 27(2):43–51.

[Johnson, 2003] Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):549–559. Publisher: Royal Society.

[Kaber and Endsley, 2004] Kaber, D. B. and Endsley, M. R. (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.

[Kahneman et al., 1982] Kahneman, D., Slovic, P., and Tversky, A. (1982). The simulation heuristic. In *Judgment under uncertainty: heuristics and biases*, pages 198–205. Cambridge University Press, Cambridge ; New York.

[Kass and Finin, 1988] Kass, R. and Finin, T. (1988). MODELING THE USER IN

NATURAL LANGUAGE SYSTEMS. *Computational Linguistics*, 14(3):18.

[Kazi et al., 2019] Kazi, S., Khaleghzadegan, S., Dinh, J. V., Shelhamer, M. J., Sapirstein, A., Goeddel, L. A., Chime, N. O., Salas, E., and Rosen, M. A. (2019). Team physiological dynamics: A critical review. *Human factors*, page 0018720819874160. ISBN: 0018-7208 Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[Kieras and Bovair, 1984] Kieras, D. E. and Bovair, S. (1984). The Role of a Mental Model in Learning to Operate a Device*. *Cognitive Science*, 8(3):255–273.

[Klein et al., 2004] Klein, G., Woods, D., Bradshaw, J., Hoffman, R., and Feltovich, P. (2004). Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intelligent Systems*, 19(06):91–95.

[Knoblich et al., 2011] Knoblich, G., Butterfill, S., and Sebanz, N. (2011). Psychological Research on Joint Action. In *Psychology of Learning and Motivation*, volume 54, pages 59–101. Elsevier.

[Kourtis et al., 2014] Kourtis, D., Knoblich, G., Woźniak, M., and Sebanz, N. (2014). Attention Allocation and Task Representation during Joint Action Planning. *Journal of Cognitive Neuroscience*, 26(10):2275–2286.

[Kourtis et al., 2013] Kourtis, D., Sebanz, N., and Knoblich, G. (2013). Predictive representation of other people's actions in joint action planning: An EEG study. *Social Neuroscience*, 8(1):31–42.

[Lausic et al., 2009] Lausic, D., Tennebaum, G., Eccles, D., Jeong, A., and Johnson, T. (2009). Intrateam Communication and Performance in Doubles Tennis. *Research Quarterly for Exercise and Sport*, 80(2):281–290.

[Le Goff et al., 2018] Le Goff, K., Rey, A., Haggard, P., Oullier, O., and Berberian, B. (2018). Agency modulates interactions with automation technologies. *Ergonomics*, 61(9):1282–1297.

[Lemoine et al., 1996] Lemoine, M. P., Debernard, S., Crevits, I., and Millot, P. (1996). Cooperation between humans and machines: first results of an experiment with a multi-level cooperative organisation in air traffic control. *Computer Supported Cooperative Work (CSCW)*, 5(2):299–321. ISBN: 1573-7551 Publisher: Springer.

[Lim and Klein, 2006] Lim, B.-C. and Klein, K. J. (2006). Team mental models and team performance: a field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior*, 27(4):403–418. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.387.

[Lipton, 1990] Lipton, P. (1990). Contrastive Explanation. *Royal Institute of Philosophy Supplement*, 27:247–266.

[Lombrozo, 2007] Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257.

[Lyons, 2013] Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series*.

[Lyons et al., 2017] Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., Smith, D., Johnson, W., and Shively, R. (2017). Shaping Trust Through Transparent Design: Theoretical and Experimental Guidelines. In Savage-Knepshield, P. and Chen, J., editors, *Advances in Human Factors in Robots and Unmanned Systems*, volume 499, pages 127–136. Springer International Publishing, Cham. Series Title: Advances in Intelligent Systems and Computing.

[Manzey et al., 2012] Manzey, D., Reichenbach, J., and Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, 6(1):57–87.

[Mar and Macrae, 2008] Mar, R. A. and Macrae, C. N. (2008). Triggering the Intentional Stance. In Bock, G. and Goode, J., editors, *Novartis Foundation Symposia*, pages 111–133. John Wiley & Sons, Ltd, Chichester, UK.

[Mathieu et al., 2000] Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., and Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of applied psychology*, 85(2):273. ISBN: 1939-1854 Publisher: American Psychological Association.

[McClure et al., 2007] McClure, J., Hilton, D. J., and Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology*, 37(5):879–901.

[McNeese et al., 2018] McNeese, N. J., Demir, M., Cooke, N. J., and Myers, C. (2018). Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 60(2):262–273.

[Mercado et al., 2016] Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., and Procci, K. (2016). Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58(3):401–415.

[Mertes and Jenney, 1974] Mertes, F. and Jenney, L. (1974). Automation Applications in an Advanced Air Traffic Management System: Volume 3. Methodology for Man-Machine Task Allocation. Technical report, United States. Dept. of Transportation. Office of the Secretary.

[Metzger and Parasuraman, 2001] Metzger, U. and Parasuraman, R. (2001). The Role of the Air Traffic Controller in Future Air Traffic Management: An Empirical Study of Active Control versus Passive Monitoring. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 43(4):519–528.

[Miller and Parasuraman, 2007] Miller, C. A. and Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human factors*, 49(1):57–75. ISBN: 0018-7208 Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[Miller, 2018] Miller, T. (2018). Contrastive Explanation: A Structural-Model Approach. *arXiv:1811.03163 [cs]*. arXiv: 1811.03163.

[Miller, 2019] Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*. arXiv: 1706.07269.

[Miller et al., 2017] Miller, T., Howe, P., and Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]*. arXiv: 1712.00547.

[Millot and Lemoine, 1998] Millot, P. and Lemoine, M. (1998). An attempt for generic concepts toward human-machine cooperation. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218)*, volume 1, pages 1044–1049, San Diego, CA, USA. IEEE.

[Millot and Pacaux-Lemoine, 2013] Millot, P. and Pacaux-Lemoine, M.-P. (2013). A Common Work Space for a mutual enrichment of Human-machine Cooperation and Team-Situation Awareness. *IFAC Proceedings Volumes*, 46(15):387–394.

[Mueller and Klein, 2011] Mueller, S. T. and Klein, G. (2011). Improving Users' Mental Models of Intelligent Software Tools. *IEEE Intelligent Systems*, 26(2):77–83.

[Nuhuis, 1999] Nuhuis, H. (1999). Automation strategies: Evaluation of some concepts. *Proceedings of the Third EUROCONTROL Human Factors Workshop Integrating Human Factors into the Life Cycle of ATM Systems*, pages 27–33.

[Ososky et al., 2012] Ososky, S., Schuster, D., Jentsch, F., Fiore, S., Shumaker, R., Lebiere, C., Kurup, U., Oh, J., and Stentz, A. (2012). The importance of shared mental models and shared situation awareness for transforming robots from tools to teammates. In *Unmanned systems technology XIV*, page 838710, Baltimore, Maryland, USA.

[O'Neill et al., 2020] O'Neill, T., McNeese, N., Barron, A., and Schelble, B. (2020). Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, page 001872082096086.

[Pacaux et al., 2011] Pacaux, M.-P., Godin, S. D., Rajaonah, B., Anceaux, F., and Vanderhaegen, F. (2011). Levels of automation and human-machine cooperation: Application to human-robot interaction. *IFAC Proceedings Volumes*, 44(1):6484–6492.

[Pacaux-Lemoine and Debernard, 2000] Pacaux-Lemoine, M. P. and Debernard, S. (2000). A common work space to support the air traffic control. *IFAC Proceedings Volumes*, 33(12):75–78. ISBN: 1474-6670 Publisher: Elsevier.

[Pacherie, 2008] Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107:179–217.

[Pacherie, 2012] Pacherie, E. (2012). The Phenomenology of Joint Action: Self-Agency versus Joint Agency. In Seemann, A., editor, *Joint Attention*. The MIT Press.

[Pacherie and Nicod, 2007] Pacherie, E. and Nicod, I. J. (2007). The Sense of Control and the Sense of Agency. *Psyche*, 13(1):1–30.

[Palmer, 1995] Palmer, E. (1995). Oops, it didn't arm-a case study of two automation surprises. In *Proceedings of the Eighth International Symposium on Aviation Psychology*, pages 227–232.

[Parasuraman et al., 2000] Parasuraman, R., Sheridan, T., and Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297.

[Perez-Osorio and Wykowska, 2020] Perez-Osorio, J. and Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, 33(3):369–395.

[Perrault et al., 2019] Perrault, R., Shoam, Y., and Brynjolfsson, E. (2019). The AI Index 2019 Annual Report. Technical report, Stanford Human-Centered Artificial Intelligence.

[Ramnani and Miall, 2004] Ramnani, N. and Miall, R. C. (2004). A system in the human brain for predicting the actions of others. *Nature Neuroscience*, 7(1):85–90.

[Rouse and Morris, 1986] Rouse, W. B. and Morris, N. M. (1986). On Looking Into the Black Box: Prospects and Limits in the Search for Mental Models. *MENTAL MODELS*, page 15.

[Sahaï et al., 2017] Sahaï, A., Pacherie, E., Grynszpan, O., and Berberian, B. (2017). Predictive Mechanisms Are Not Involved the Same Way during Human-Human vs. Human-Machine Interactions: A Review. *Frontiers in Neurorobotics*, 11:52.

[Salas et al., 1995] Salas, E., Prince, C., Baker, D. P., and Shrestha, L. (1995). Situation Awareness in Team Performance: Implications for Measurement and Training. *HUMAN FACTORS*, page 14.

[Sarter and Woods, 1994] Sarter, N. B. and Woods, D. D. (1994). Pilot interaction with cockpit automation II: An experimental study of pilots' model and awareness of the flight management system. *The International Journal of Aviation Psychology*, 4(1):1–28. ISBN: 1050-8414 Publisher: Taylor & Francis.

[Sarter and Woods, 1995] Sarter, N. B. and Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human factors*, 37(1):5–19. ISBN: 0018-7208 Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[Scharre and Horowitz, 2015] Scharre, P. and Horowitz, M. C. (2015). An Introduction to Autonomy in Weapon Systems. Technical report, Center for a new American Security.

[Sebanz et al., 2006] Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76.

[Sebanz et al., 2003] Sebanz, N., Knoblich, G., and Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition*, 88(3):B11–B21.

[Sebanz et al., 2005] Sebanz, N., Knoblich, G., and Prinz, W. (2005). How two share a task: corepresenting stimulus-response mappings. *Journal of Experimental Psychology: Human Perception and Performance*, 31(6):1234. ISBN: 1939-1277 Publisher: American Psychological Association.

[Shea et al., 2014] Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., and Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, 18(4):186–193.

[Sheridan and Verplank, 1978] Sheridan, T. and Verplank, W. (1978). Human and computer control of undersea teleoperators. *Human and Computer Control of Undersea Teleoperators*.

[Sheridan, 1987] Sheridan, T. B. (1987). Handbook of human factors. In *Supervisory Control*, pages 1243–1268. Wiley-Interscience.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489. Citation Key Alias: silverMasteringGameGo2016.

[Simon, 1990] Simon, J. R. (1990). The Effects of an Irrelevant Directional CUE on Human Information Processing. In Proctor, R. W. and Reeve, T. G., editors, *Advances in Psychology*, volume 65 of *Stimulus-Response Compatibility*, pages 31–86. North-Holland.

[Stanton, 2016] Stanton, N. A. (2016). *Distributed situation awareness*, volume 17. Taylor & Francis. Issue: 1 Pages: 1-7 Publication Title: Theoretical Issues in Ergonomics Science.

[Stanton et al., 2011] Stanton, N. A., Dunoyer, A., and Leatherland, A. (2011). Detection of new in-path targets by drivers using Stop & Go Adaptive Cruise Control. *Applied Ergonomics*, 42(4):592–601.

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: an introduction*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., nachdr. edition. OCLC: 837901590 Citation Key Alias: suttonReinforcementLearningIntroduction20, suttonReinforcementLearningIntroduction20a, suttonReinforcementLearningIntroduction20b.

[Swain and Guttmann, 1983] Swain, A. D. and Guttmann, H. E. (1983). Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report. Technical Report NUREG/CR-1278, SAND-80-0200, 5752058, US NUCLEAR REGULATORY COMMISSION.

[Tsai et al., 2008] Tsai, C.-C., Kuo, W.-J., Hung, D. L., and Tzeng, O. J. L. (2008). Action co-representation is tuned to other humans. *Journal of Cognitive Neuro-*

science, 20(11):2015–2024.

[Van Charante et al., 1992] Van Charante, E. M., Cook, R. I., Woods, D. D., Yue, L., and Howie, M. B. (1992). Human-computer interaction in context: Physician interaction with automated intravenous controllers in the heart room. *IFAC Proceedings Volumes*, 25(9):263–274. ISBN: 1474-6670 Publisher: Elsevier.

[Vantrepotte et al., 2022] Vantrepotte, Q., Berberian, B., Pagliari, M., and Chambon, V. (2022). Leveraging human agency to improve confidence and acceptability in human-machine interactions. *Cognition*, 222:105020.

[Vezhnevets et al., 2017] Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., and Kavukcuoglu, K. (2017). FeUdal Networks for Hierarchical Reinforcement Learning. *arXiv:1703.01161 [cs]*. arXiv: 1703.01161.

[Wiener, 1989] Wiener, E. L. (1989). Human factors of advanced technology (glass cockpit) transport aircraft. Technical report, NASA Ames Research Center.

[Wilson and Rutherford, 1989] Wilson, J. R. and Rutherford, A. (1989). Mental Models: Theory and Application in Human Factors. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 31(6):617–634.

[Woods et al., 2004] Woods, D., Tittle, J., Feil, M., and Roesler, A. (2004). Envisioning human-robot coordination in future operations. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):210–218. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).

[Woods and Tinapple, 1999] Woods, D. D. and Tinapple, D. (1999). W3: Watching human factors watch people at work. In *Presidential address, presented at the 43rd Annual Meeting of the Human Factors and Ergonomics Society, Houston, TX. Multimedia production available at http://csel. eng. ohiostate. edu/hf99/. Announcement deadlines: 1st day of the month prior to the desired issue.*

[Yoshida et al., 2008] Yoshida, W., Dolan, R. J., and Friston, K. J. (2008). Game theory of mind. *PLoS Comput Biol*, 4(12):e1000254. ISBN: 1553-7358 Publisher: Public Library of Science.

[Yoshida et al., 2010] Yoshida, W., Seymour, B., Friston, K. J., and Dolan, R. J. (2010). Neural Mechanisms of Belief Inference during Cooperative Games. *The Journal of Neuroscience*, page 8.