Interdependence and trust analysis (ITA)

a framework for human–machine team design

Centeio Jorge, Carolina; Jonker, Catholijn M.; Tielman, Myrthe L.

Interdependence and trust analysis (ITA): a framework for human–machine team design

Carolina Centeio Jorge, Catholijn M. Jonker & Myrthe L. Tielman

Published online: 25 Nov 2024.

Submit your article to this journal ⬈

Article views: 197

View related articles ⬈

View Crossmark data ⬈

Taylor & Francis
Taylor & Francis Group

Interdependence and trust analysis (ITA): a framework for human–machine team design

Carolina Centeio Jorge[a], Catholijn M. Jonker[a,b] and Myrthe L. Tielman[a]

[a]Intelligent Systems, Delft University of Technology, Delft, The Netherlands; [b]LIACS, Leiden University, Leiden, The Netherlands

ABSTRACT

As machines' autonomy increases, the possibilities for collaboration between a human and a machine also increase. In particular, tasks may be performed with varying levels of interdependence, i.e. from independent to joint actions. The feasibility of each type of interdependence depends on factors that contribute to contextual trustworthiness, such as team members' competence, willingness and external factors. In this paper, we present the Interdependence and Trust Analysis (ITA) framework, which is an extension of Coactive Design's Interdependence Analysis framework (Johnson, M., J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. Birna Van Riemsdijk, M. Sierhuis. 2014. Coactive Design: Designing Support for Interdependence in Joint Activity. *Journal of Human-Robot Interaction* 3 (1): 43–69. https://doi.org/10.5898/JHRI.3.1.Johnson). By including information on contextual trustworthiness, ITA can better support the design of human–machine teams, as well as task allocation and selection. Evaluated through expert interviews and a focus group involving a search and rescue scenario, ITA shows potential as a decision-making tool and a communication bridge among human and machine teammates. Our findings emphasise the need to define tasks and roles based on agent characteristics, and imply that decision-making models should align with human-centred objectives. ITA also highlights the trade-off between utility and effort when designing trustworthy systems, suggesting that guided conversations could improve the team design process. Finally, the ITA framework may improve transparency, justification, and interpretability in decision-making, contributing to appropriate trust among teammates.

1. Introduction

In scenarios where humans and machines collaborate, several design decisions have to be made, such as who does what (Ali et al. 2022; Azevedo-Sa, Yang, et al. 2021). In some situations this may be straightforward, such as when there is no overlap of teammates' (human's or machine's) expertise, e.g. imagine a kitchen robot that only works as a pressure cooker and a human (who, of course, cannot work as pressure cooker) who can prepare the ingredients that go in the machine. On the other hand, there are situations where both teammates can do certain tasks, for example a kitchen robot arm that also chops vegetables and a person who can do the same. Situations like the latter may become more frequent with the advancement of AI, since machines have the possibility of functioning with higher levels of autonomy. This opens the door for interdependence between humans and machines, i.e. when two parties have to rely on each other to perform a joint activity (Johnson et al. 2014). Growing capabilities and autonomy mean more possible designs for human–machine collaborations, with different types of interdependence in the subtasks involved, from full independence to mandatory joint actions. Finding a good division of labour between machine with variable levels of autonomy and human teammates is a problem (Seeber et al. 2020) in human–machine team design, which can be helped with methodological interdependence and trust analysis, as proposed in this paper.

The design of human–machine interdependent relationships for teamwork involves a symbiosis between humans and AI that benefits humans (Van Zoelen et al. 2023). In Johnson et al. (2014), the authors present the Interdependence Analysis table as a framework for Coactive Design (Johnson et al. 2014), i.e. an approach to addressing the increasingly sophisticated roles that people and machines play as the use of human–machine teamwork expands into new, complex domains. The original Interdependence Analysis lists the capacities required for the execution of tasks. It encourages a comprehensive analysis of which

teammate has the required capacities to be a performer of a certain task and whether the other teammate's support is mandatory, not possible, or helpful, i.e. increasing reliability or efficiency. After filling in the table, one should be able to understand the necessary interdependencies for each task, through a colour code and requirement gathering.

Although a thorough analysis of capacities is an important step, we claim that it is also important to consider other dimensions that may lead to the success of a task. Models based on trust and trustworthiness between humans (human–human) have been developed to formalise the dimensions that may lead artificial agents to successfully perform tasks (Falcone and Castelfranchi 2004; Falcone et al. 2013). Following these models, for a cognitive agent, either human or artificial, to successfully perform a task, they need to have the capacities/capabilities (i.e. 'can they do it?'), the willingness/intention to do it (i.e. 'will they do it?'), and to have the external opportunities/permissions to do it (i.e. 'is it possible to do it?'). In other words, these dimensions can be used to assess the trust that agents have in their teammate(s) to successfully perform a certain task. Trust in human–machine teams includes natural trust, i.e. trust beliefs of the human (see e.g. Vinanzi, Cangelosi, and Goerick 2021), and artificial trust, i.e. trust beliefs of the machine (see e.g. Centeio Jorge, Jonker, and Tielman 2024b). What makes a human trustworthy for a task is not necessarily what makes a machine trustworthy for that task (Ulfert et al. 2023), however, to decide who should do it, we need to consider both. So far, there is no framework supporting human–machine team design that considers, in a methodological way, both machine and human team members' contextual trustworthiness (not only capabilities but also willingness and external factors) for different interdependent roles and tasks. In Johnson and Bradshaw (2021), Johnson et al. already propose an extension of the table that includes trust as one extra dimension to consider when analysing interdependencies. However, this dimension of trust is (1) only considered for the trust in one of the agents (the performer) involved in the interdependence, and, in our opinion, it (2) could also be further divided into dimensions that are easier to assess, update and use for informed decision-making. As such, we propose to include an analysis of teammates' willingness and external factors regarding different team configurations in the process of human–machine team design.

This paper's contribution is the *Interdependence and Trust Analysis (ITA)* framework, centred on an extended new version of Johnson's Interdependence Analysis tables, the *ITA table*. The ITA framework presents the conceptual workflow of the dynamic information used for decision-making in human–machine teams, which serves as input and output for the ITA table. Additionally, the ITA table analyses three dimensions of a team member's trustworthiness, i.e. not only *competence* (as in the original table in Johnson et al. 2014), but also *willingness* and *external factors*, for the different tasks involved in a human–machine shared goal. Our method proposes that human–machine team design should consider willingness as an important dimension in assessing the feasibility of a team configuration in terms of interdependence and task allocation. This implies that we conceptualise all team members (machines and humans) as agents with intentions (i.e. *something that [one] wants and plans to do* as per the Cambridge Dictionary). They can not only act, but also choose which possible actions to do. This is in line with frameworks like (Falcone and Castelfranchi 2004; Georgeff et al. 1998), used in multiagent systems, but goes beyond the traditional view of machines as just executioners of an action when interacting with humans. Although this intentionality is widely accepted for human teammates, there is still a tendency to overlook willingness even for human team task allocation, and most works consider only capabilities, see e.g. Saad, Hindriks, and Neerincx (2018), Ali et al. (2022) and Johnson et al. (2014). Furthermore, the willingness dimension should be considered a task-based and role-based characteristic, rather than a property of the teammate that is transversal to all tasks and interdependencies. For example, a human teammate may be willing to independently carry a light object but not willing to carry it together with a robot. However, they may be willing to carry an object together with a robot if the object is heavy. This implies that human–machine team design should also consider that willingness depends on roles, for example, allocating the task of carrying light objects to the human while having the robot assist could decrease team performance and the overall human experience. This is in line with (Noormohammadi-Asl et al. 2022, 2023), who suggests that a machine should adjust to the human's preference of being a leader or a follower on collaborative tasks. However, these works overlook joint actions and the possibility that capabilities for each role may also differ (e.g. one may not have the strength to carry a heavy object alone, but has some strength to support another teammate carrying it), which we include in our framework. Additionally, we suggest that external factors are increasingly relevant to consider in human–machine teamwork design, as machines become more autonomous and require clearer boundaries from performing certain

actions, such as ethical and moral decisions (e.g. deciding whether to save someone's life), or for safety measures (e.g. holding a gun). Some works defend the development of artificial moral agents, i.e. artificial agents capable of making ethical and moral decisions (Cervantes et al. 2020), while others defend Meaningful Human Control (MHC), i.e. humans should ultimately remain in control of, and thus morally responsible for, everyday actions (Santoni de Sio and van den Hoven 2018). This implies that our human–machine team design allows the human to explicitly delimit the machine's permissions and detect situations that require human oversight and control (aligned with van der Waa et al. 2021). Finally, our framework implies that human–machine team design decisions need to be explicit and easily revisited in order to comply with new ethical guidelines (e.g. transparency and traceability), such as *European AI Act*, and the *IEEE Ethically Aligned Design*. To evaluate the ITA table, we conducted two dyadic interviews (with two participants each) and one expert focus group with five participants. We present the results through a thematic analysis.

The proposed Interdependence and Trust Analysis framework can be used by a team designer to make decisions regarding which role each teammate should have in different tasks. Furthermore, it could be used as a decision-making support system, as well as a shared mental model (Gervits et al. 2020; Salas, Sims, and Burke 2005; van de Kieft, Jonker, and Birna van Riemsdijk 2011). Using the analysis for such cases may increase transparency among teammates and facilitate justification of one's actions, and, consequently, appropriate trust (Ulfert et al. 2023; Verhagen et al. 2022).

In Section 2 we start by presenting the background concepts and related work that sustain our work. Then, in Section 3 we present the table, and frame it in Section 4. We present the results of the evaluation of the table in Section 5 and 6 and discuss it in Section 7.

2. Interdependence and trust analysis (ITA)

Human–machine (and human-AI, human-agent, etc) teamwork studies aim at integrating humans and intelligent machines, rather than deliberately pushing the human out of the loop (Sierhuis et al. 2003). The goal is usually to provide support to the human, avoiding hazardous consequences (Gervits et al. 2020). In fact, these teams can be beneficial for humans, for example in situations where it can be unsafe to have humans doing everything, e.g. disaster response (De Greeff et al. 2018) and search and rescue (Saad, Hindriks, and Neerincx 2018). In other cases, these teams can reduce the human's workload, e.g. in collaborative

cooking (Goubard and Demiris 2023) and collaborative driving (Azevedo-Sa, Jayaraman, et al. 2021) scenarios. These teams can also be effective for tasks that require high precision, e.g. robot-assisted surgeries (Cypko et al. 2022). However, the design and implementation of these teams pose challenges (Klien et al. 2004; van den Bosch et al. 2019), especially when machines start having more autonomy as their range of capabilities increases. More autonomy allows for different possibilities of *interdependence*, depending on the scenario, which may allow for different team designs (who does what, etc). Furthermore, there are moments when machines should not use their capabilities, in order to comply with social norms, and ethical principles (Baum et al. 2023), and always allowing for meaningful human control (van der Waa et al. 2021).

Team members need to cooperate, collaborate and coordinate (Johnson et al. 2014). This is only possible with communication, mutual trust and shared mental models (Salas, Sims, and Burke 2005). Designing human–machine teams should ensure these mechanisms, which can be challenging. In particular, finding a good division of labour between machine and human teammates is one such challenge (Seeber et al. 2020). In the process of task selection or allocation (see e.g. Abuhaimed, Karaoglu, and Sen 2023; Noormohammadi-Asl et al. 2022), a team member or designer, respectively, needs to consider how much they trust different team members to successfully perform a task within a certain context (Ali et al. 2022). In the context of human–machine teamwork, we see trust as the belief in an entity's trustworthiness to perform a task successfully, within a certain context (Centeio Jorge, Jonker, and Tielman 2023). Trustworthiness is a complex concept, and following the literature, it can consist of a set of dimensions that range from the trustee's competence to its intentions (De Visser et al. 2020; Griffiths 2005). Depending on the nature of the trustor and trustee, the trust and trustworthiness constructs may be more or less adequate. There are several works studying how humans trust machines (see e.g. Law and Scheutz 2021; Lee and See 2004; Lee and Sun 2023; Rezaei Khavas et al. 2024), but not so many showing how machines should trust human partners (see e.g. Vinanzi, Cangelosi, and Goerick 2021; Vinanzi et al. 2019). Models in slightly different settings propose that trustworthiness depends on (1) Ability, Benevolence and Integrity (Mayer, Davis, and Schoorman 1995) (in human organisations), (2) Willingness, Competence (Castelfranchi and Falcone 2010) (in multi-agent systems) and (3) Performance, Process and Purpose (Lee and See 2004) (when the human is the trustor and artificial agent is the trustee). For this last case, Hancock

et al. (2011) proposes that the agent's characteristics affecting trust (i.e. perceived trustworthiness) are performance-based (such as reliability, failure rate, etc) and attribute-based (such as anthropomorphism, robot personality, etc).

Although there are several interpretations about what exactly trustworthiness is, we see a tendency to separate it into two bigger dimensions, i.e. one related to the potential to execute a task successfully (e.g. ability, competence, performance), and another related to the behaviour that may influence the execution of the task, related to the factors that contribute to one's intention of performing a task (e.g. benevolence, integrity, willingness, process, purpose). In fact, these two main dimensions are used in works such as Ullman and Malle (2020), where authors divide human trust into performance trust and moral trust. Similarly, McKee, Bai, and Fiske (2022) shows how humans, besides competence, also perceive warmth in artificial teammates, which also affects their decision-making and collaboration (Centeio Jorge, Jonker, and Tielman 2024b). In summary, to assess an agent's trustworthiness to successfully execute a certain task, we need to take into account the agent's competence and willingness (following Falcone and Castelfranchi 2004's nomenclature) for the execution of that task.

Furthermore, the COM-B model for behaviour change (Michie, Van Stralen, and West 2011) suggests that besides capability and motivation, which align with the two trustworthiness dimensions explored in the previous paragraph, a person needs the opportunity to behave in a certain way. In the context of teamwork, opportunity is only possible when a task is available and possible (Bradshaw et al. 2004). In other words, the execution of a task is influenced by external factors, which are contextual conditions determining the situation in which the task is executed (Falcone et al. 2013; Hancock et al. 2011), such as team setting, environmental configuration, emotional state, workload, etc. As such, to entrust a task to an agent, one needs to have e positive belief regarding the agent's trustworthiness (i.e. competence and willingness), as well as a positive belief that the external factors allow that agent to execute that task.

When collaborating, humans and machines can take different roles (van Diggelen and Johnson 2019), i.e. they can have different interdependent relationships. Interdependence can be soft or hard (Johnson et al. 2014). Soft interdependence happens when the collaboration improves the task efficiency, but it is not required. On the other hand, hard interdependence happens when the collaboration is necessary for the task to be successful. In particular, in soft interdependencies there can be a performer and a supporter (Johnson and Bradshaw 2021). The supporter, a teammate that possibly (necessarily or not) helps the performer, the main teammate involved in completing a task, to do the task. As such, when designing a human–machine team, in particular, deciding how to select or allocate tasks, one can select or allocate a specific role to perform a task. What's more, the competence and willingness required to be a main performer may differ from those of being a supporter (Noormohammadi-Asl et al. 2022). Particularly, human teammates may have more or less willingness to engage in interdependent relationships with the machine, depending on the human characteristics or machine's characteristics. Human factors that contribute to one's attitude towards a machine are related to the personal discomfort and concerns in various interaction scenarios (Nenna et al. 2024), which tend to affect the human trust in the machine. On the other hand, machine's characteristics that may affect the human's willingness to collaborate range from machine's appearance (see e.g. Song and Luximon 2024) to machine's previous behaviour, such as failure history (Centeio Jorge et al. 2023). As such, distinguishing the levels of competence and willingness for the different interdependencies gives more insight about the different feasible team configurations.

In this paper, we aim at providing a structured analysis of the dimensions of a performer's competence, willingness and external factors and evaluate the feasibility of each possible interdependence relationship. The final decision of which interdependence is better for a certain task is left to the user (and trustor) to decide, as this mainly depends on their perceived risk (Fahnenstich, Rieger, and Roesler 2023; Hoesterey and Onnasch 2023; Stuck, Holthausen, and Walker 2021; Stuck, Tomlinson, and Walker 2022; Wagner, Robinette, and Howard 2018), of trusting and, sometimes, of not trusting, see e.g. Mehrotra et al. (2024). This is related to the formal belief of dependence, as in Falcone and Castelfranchi (2004), which is how much an agent believes they depend on another entity for a certain goal.

3. Table for ITA

The goal of our proposed analysis through a table is twofold. Firstly, we want a framework that supports team design by providing a more comprehensive analysis of all possible team configurations based on the feasibility of the interdependencies at the atomic task (i.e. a task that is not composed of subtasks) level. Moreover, for each of these interdependencies, we want a framework that analyses the trustworthiness of each teammate for a certain role. For this, we analyse not only the competence/performance dimension, but also the

Atomic Task	Possible Performer(s)	Can? (skills, knowledge)	Will? (intention, preference)	Ext. Factors (opport., resources)	F	Configuration Feasibility					Design choice
						H	H+	H+M	M+	M	
Washing	Joint	✓	✓	✓	✓						machine performer + human support
	H independent	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	M independent	✓	✓	✓	✓						
Peeling	Joint	✓	✓	✓	✓						mandatory joint
	H independent	✓	X	✓	X	X	X	✓	✓	✓	
	M independent	✓	✓	✓	✓						
Chopping	Joint	✓	✓	X	X						human performer + no support
	H independent	✓	✓	✓	✓	✓	X	X	X	X	
	M independent	✓	✓	X	X						
Frying	Joint	X	X	✓	X						machine performer + no support
	H independent	X	X	✓	X	X	X	X	X	✓	
	M independent	✓	✓	✓	✓						

Figure 1. Interdependence and trust analysis table for several atomic tasks that compose the major task of *making fries*. Possible performers are the human (H), the machine (M) or both being co-performers (Joint). To each of these possibilities, we analyse whether they have the skills and knowledge to do a task (whether that performer *can*), whether they have the intention and preference to do the task (the performer *will*) and, finally, if the external factors and permissions allow. We can see the resulting feasibility (F) of each performer option and the resulting feasible configurations that leads to. The last column presents the design choice. The areas that are within the dark red scattered lines are the ones that can be altered by the user. Column *F* and *Configuration Feasibility* are automatically calculated.

willingness/intention dimension, as well as the external factors that may restrain that action. This explicit information should improve the process of design and decision-making in human–machine teams for task selection and/or allocation, whether this is to be done by a team member or for a team designer (i.e. not necessarily involved in performing the tasks). The table can be found in Figure 1. The parts of the table surrounded by scattered thick dark-red line are to be filled in by the users. Besides those, the table is automatically filled in. In this section, we present the structure of the table and how it can be used.

3.1. Structure of the table

3.1.1. Atomic tasks
When there is a team goal, this needs to be divided into sub-tasks, which in turn can be divided into other sub-tasks, repeatedly, until the goal is divided into atomic tasks. We call atomic tasks the tasks that do not need to be broken down into smaller tasks. The analysis of interdependence and trust will focus on each of these tasks individually. They are to be decided by the user (further explained in Section 3.2.2).

3.1.2. Possible performer(s)
For each atomic task, we need to consider who can perform it. The possibilities of performing an action are doing it independently, i.e. the human as performer (H independent), or the machine as performer (M

independent) or doing it jointly (Joint), as a hard interdependence. Each of these three potential performers will be analysed in terms of dimensions of trustworthiness, for each task.

3.1.3. Dimensions
Based on the literature presented in Section 2, we include (1) a belief related to ability, performance, competence (column *can* in Figure 1), (2) a belief which comprehends everything besides ability that may contribute to the choice of performing a task successfully, i.e. willingness, benevolence, integrity and personal preference/motivation for a certain task and (3) the context which comprehends external factors (opportunities, permissions), in our analysis. We can find the columns *Can?*, *Will?* and *Ext. Factors* on the table.

3.2. How to use the table

3.2.1. Scenario
The scenario that was used to fill in this table was inspired by cooking scenarios (used in human–robot interaction studies Goubard and Demiris 2023 and human-AI collaboration studies, such as the test bed Overcooked-AI Carroll et al. (2019)) and consisted of *making fries* with a set of constraints. The constraints were:

- The machine is not allowed to hold a knife (which impedes chopping).

- The human does not want to fry the potatoes, because they are afraid of getting burnt.
- The human does not know how to fry potatoes.
- The human does not want to peel potatoes if they are the only one doing it.

3.2.2. Step 1: atomic tasks

The first step for the interdependence and trust analysis is to know what tasks need to be done. As such, users of the table must first agree on which atomic tasks need to be listed in the table. Determining the atomic tasks and their level of detail can be challenging, depending on the scenario. However, the idea is to divide the tasks until the point when it is clear what *Joint*, *Independent H* and *Independent M* may look like, so that we can assess the possible performers' trustworthiness. We decided that *making fries* includes the atomic tasks of washing, peeling, chopping and frying. After establishing the atomic tasks, the user should start filling in the areas of the table that are surrounded by scattered thick dark-red line (in Figure 1). In particular, the user should fill in the atomic tasks on the table, in the first column.

3.2.3. Step 2: assessing trustworthiness

The second step of the ITA analysis is to assess the trustworthiness of the different possible performers, for each task, by signing each dimension with a '✓' if positive or with 'X' if negative. For example, when we analyse whether the human can perform the task independently, we should consider whether they can (i.e. have the competences, skills, knowledge…), whether they will (i.e. want to, would choose to do that task) and, finally, if they have the external opportunities and resources to do it (i.e. external factors).

For example, the machine was not allowed to hold a knife, which should make the cells of *external factors* negative (X) both for *M (machine) independent* and *joint* in the chopping task. Also, the human did not want to fry, because they were afraid to do it, but also did not know how to. This information should make *Can?* and *Will?* negative (X) for *H (human) independent* and *Joint* in *Frying*. Finally, the human did not want to peel potatoes alone, which puts an X on *Will?* for *H (human) independent*.

3.2.4. Step 3: interpret feasibility columns

3.2.4.1. Performer Feasibility (F) column. After filling in the table with the trustworthiness information, the column *F* will present the feasibility of each performer, for each atomic task. This feasibility is negative (X) if at least one of the dimensions is not feasible (i.e. there is an X in one of the dimensions) and positive (✓) otherwise. With the information of which performers are

possible for each atomic task, we can infer which configurations are feasible.

3.2.4.2. Configuration Feasibility column. The team configurations are the combinations of possible roles that each team member can take for a certain task, i.e. the different interdependencies that can happen in a task. We consider five possible team configurations (under *Configuration Feasibility* header) for a team composed of one human and one machine. If we consider independent configurations, we can have either a completely independent human performer (H), or a completely independent machine performer (M). There are also two possible soft interdependencies, i.e. human with support (H+), which happens when the human can be independent, but support is possible to increase efficiency or reliability, and machine performer with human support (M+). Finally, there is also a hard interdependence, i.e. mandatory joint (H+M), where human and machine have to co-perform the task. The configurations' feasibilities are inferred from the performers' feasibilities (see Figure 2). For example, we consider that if *H independent* is feasible, then the team configuration *human performer + no support (H)* is also feasible. Having a feasible joint performer also leads to a feasible *mandatory joint (H+M)* configuration. We infer the supporting roles given that joint is possible, i.e. if joint is possible, support is also possible (see the *peeling* example). In the ITA table (in Figure 1), we can see for each task which configurations are feasible. For example, for *washing*, all configurations are feasible

Team Configuration	Joint	H ind.	M ind.
human performer + no support (H)		✓	
human performer + machine support (H+)	✓	✓	
mandatory joint (H+M)	✓		
machine performer + human support (M+)	✓		✓
machine performer + no support (M)			✓

Figure 2. For each team configuration to be considered feasible, a set of performers need to be feasible as well. This table shows which performers need to be feasible (in F column) for a team configuration to be considered feasible (in Configuration Feasibility column).

whereas for *chopping* only human performer without support seems feasible.

3.2.5. Step 4: design choice

Once the user knows what the feasible configurations are, they have the basic information to make a decision. In the *Design Choice* column, the user can pick one of the configurations for each atomic task. The table does not advise for any design. Depending on the tasks and scenarios, we believe there will be other things to consider when deciding which of the feasible configurations to pick (e.g. workload, values, time to finish task). Although that is out of the scope of this paper, we discuss it further in Section 7.1.5.

4. Framework

The interdependence and trust analysis (ITA) framework is the conceptual workflow of the dynamic information used for decision-making in human–machine teams, which serves as input and output for the ITA table. The ITA framework can not only be used by a team designer, with an overview of all tasks and teammates, but also by the teammates themselves, either human or artificial. We envision the ITA framework to be used in two main ways, both having similar but potentially slightly different requirements on the ITA table:

(1) *ITA for human use:* A table for human teammates or human team designers to use, which includes the assessment of different trustworthiness dimensions (competence and willingness), and of one context (external factors) dimension, for different team configurations.

(2) *ITA for artificial use:* A computed version of the table (with the same dimensions) which can be used by an artificial teammate or artificial team designer.

Independently of being used by a human or artificially, the framework that surrounds the table is conceptually the same. Figure 3 presents the framework, and the modules that treat the information that goes in and out of the table, in the process of design and decision-making. The two main modules that need to be specified when used the table are *Information Collection* and *Access*. How the information that goes in the table is collected is entirely dependent on the table's use and user. For example, if the table is being used by an artificial agent, the information collection can either be done through sensors and/or machine learning models, or inputted directly by a human. Similarly, who has access to which table is decided by the team designer. It can be that all team members have a table of their own and can also see others' tables, or just one person has a table, for example the team designer, and this table is private.

Furthermore, this table can be attached to other modules, to be decided by the user/developer. In particular, how the information in the table is communicated to other team members is to be decided by the user, e.g. the user can have explanations generated from the values of the table. Similarly, the decisions that derive from this analysis, how this information is processed, and how it is applied, are also up to the user, e.g. there may be a module that uses the table for task allocation. Finally, the consequences of the applications of the table should lead to updating
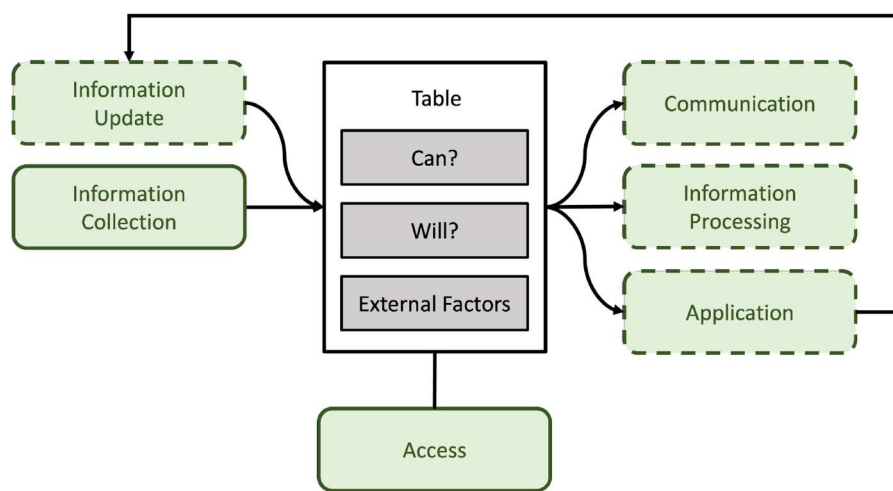
Figure 3. The process of decision-making with the help of the table requires information collection (and updates), defined access permissions, and the processing, application and communication of the structured information of the table. All these modules are entirely dependent on the use of the table and also on what the user/designer prefers.

the table's information. This module can be related to how information is collected, but can also be something different entirely. For example, the first time information is collected it may be from a human source (e.g, a manager), but during teamwork sessions (or after), this information can be updated automatically by an algorithm.

There are several potential uses of the ITA table and framework, depending on which modules the users wish to add. Primarily, the Interdependence and Trust Analysis framework is suitable for task selection and allocation. For example, a machine can compute the table and use it to make decisions on whether to support the human or not, or who to call for help for a certain task (see more in Centeio Jorge, Jonker, and Tielman 2023). If it is possible for the machine to update its beliefs regarding other teammates and itself according to this structure and representation, this framework can provide transparency and potentiate justifications from the artificial teammate. For instance, the machine can explain that it decided to fry veggies, because it believes that its human teammate is not willing or capable to do it. This can potentially happen either by presenting the table itself or generating text from it.

In fact, the table can also be seen as a formalisation of the information collected by team members and team designers, and it can provide shared mental models and communication (as per Salas, Sims, and Burke 2005). For example, if teammates can share each other's tables with each other, or have a centralised one (at least for certain dimensions), it is possible to see when beliefs are misaligned. To illustrate, perhaps I believe that the external factors do not allow a certain performer to execute a task, but my teammate disagrees. This can be perceived through sharing table information among team members. This being said, this framework also offers a good analysis of dyadic (and possibly team) trust, which can facilitate appropriate and warranted trust among teammates (Lewis, Li, and Sycara 2021) by guaranteeing that the teammates' beliefs are aligned.

5. Evaluation

We split the evaluation of the table in two phases. The first phase was composed of two expert interviews with two participants each (dyadic interviews) and was intended to improve the table. The second phase was one focus group composed of five participants, and was meant to evaluate the final version of the table (the one presented in this paper). Before conducting the experiments, we obtained approval from the ethics team of Delft University of Technology (ID nr 3488).

5.1. First phase of evaluation

Dyadic interviews provide several advantages, such as allowing the interviewer to observe deeper discussions than in an individual interview (Morgan et al. 2013; Szulc and King 2022). At the same time, it is easier to find available and compatible pairs than groups, which brings an advantage when comparing to focus groups. We ran two dyadic interviews in person, which lasted one hour and a half each. They were composed of (1) analysing interdependence and trust of a collaborative task (to be executed by a team composed of one human and one machine) by filling in our proposed table, and (2) answering six open questions. The presented table was an earlier, extended version of the one presented in this paper (in Appendix A.1). It included the thorough analysis of all dimensions for all five possible interdependence configurations. Furthermore, instead of checkmarks, that version made use of a colour code for feasibility, and the columns had a slightly different name, while referring to the same concepts.

5.1.1. Participants

Each dyadic interview of the first phase of evaluation was composed of two experts who were researchers in the field of human–machine interaction and collaboration. They were two men and two women (one man and one woman in each group), with ages between 25 and 35.

5.1.2. Task

The scenario presented to these two dyadic interviews was very similar to the one presented in Section 3.2.1. However, instead of making fries, participants were told the scenario was about frying veggies, which did not include the peeling task (as in Appendix A.1). The set of constraints were

> 'Both machine (M) and human (H) can wash and are willing to do it. However, the human is not willing to support, though, as she thinks it is not necessary. For safety reasons, the machine should not use the knife. Finally, the human does not know how to fry, and she is scared of it too, but can help, and the machine can only do it with help of others.'

Participants were explained the dimensions and interdependence configurations included in the table and asked to fill it in together, without being presented with an example beforehand. Furthermore, the questions (inspired by Krueger and Casey 2002) we asked participants at the end included 'What is the one thing you liked best/least?', 'What would you change/keep in the table?' and 'In which situations would you use/not use the table?'.

5.1.3. Data processing

This first phase of evaluation served as a pilot and no structured analysis was made. The authors went through the experts' comments during the tasks and their answers to the open questions, and summarised the most predominant comments. These comments were then used to improve the table.

5.2. Second phase of evaluation

After the first phase of evaluation, the table was changed according to the feedback received, taking the shape that we present in this paper. The second part was aimed at evaluating the current table's final usability with a use case in the domain of firefighting. We opted to do this evaluation online, through MS Teams, since (1) we included participants from different physical locations and (2) it was easier to collect and process the transcripts. The session lasted one hour and a half.

5.2.1. Participants

This focus group counted on five participants, three men and two women, with ages between 25 and 55. Two of the participants were firefighters, and the other three were researchers in the field of human-agent teamwork (applied to the fields of firefighting, military and manufacturing), with backgrounds in Psychology and Computer Science.

5.2.2. Use case

For a more realistic evaluation of the table, we looked for a scenario where a human–machine team is currently developing. As such, two of the participants worked in a fire department which has been moving towards more autonomous solutions in recent years. In particular, they have a robot, which we will call *Rob* for simplification, which is capable of moving, recording in real time, extinguishing fires, among other things. Rob is currently controlled by another firefighter through a tablet. The use case in this focus group was based on the possibility of having Rob moving autonomously. We believe that people that are already dealing with the challenges of such teams can give better insight regarding the usability of our table, including the positive and negative aspects of it.

5.2.3. Task

Participants started by being presented to the main concept of interdependence and the different interdependence configurations in a human-agent team. After this, we presented our table pre-filled with the cooking example, which was presented in the first phase as the main activity. Finally, participants were given the use case and twenty minutes to complete the table together, regarding the presented use case. They were asked to think out loud.

In particular, the participants were told

'*Let's say that we have the situation of a building with fire, and you need, as a team, to locate people inside the building. So the subtasks of this task are moving in general, which is composed of choosing where to move, i.e. **planning the trajectory**, and also the actual **movement**; and clearing the spaces, i.e. **scanning/observing** and **processing what is scanned/observed**. Imagine that there is a team composed of a firefighter and Rob, the robot. The environment does not allow the human firefighters to go in, you can imagine that it can be for several reasons. Imagine also that Rob can go in and has autonomy. In particular, Rob can move autonomously, but it can also be teleoperated (i.e. the human choses the trajectory). It can also scan the environment around and provide some analysis into the scans. However, the scans should be checked by the human firefighter as well.*'

The task and subtasks can be found in Figure 4.

After twenty minutes, the participants were asked the following questions (inspired by Krueger and Casey 2002):

Q1 What one thing do you like the best?
Q2 What one thing do you like the least?
Q3 Under what circumstances would you use the table?
Q4 Under what circumstances would you not use the table?

5.2.4. Data processing

To analyse this second phase of evaluation, we collected transcripts and ran a thematic analysis (Braun and Clarke 2006). The transcripts were divided into five parts, each originating a different coding scheme. We divided the transcripts collected during the activity, and then for each question, Q1–Q4.
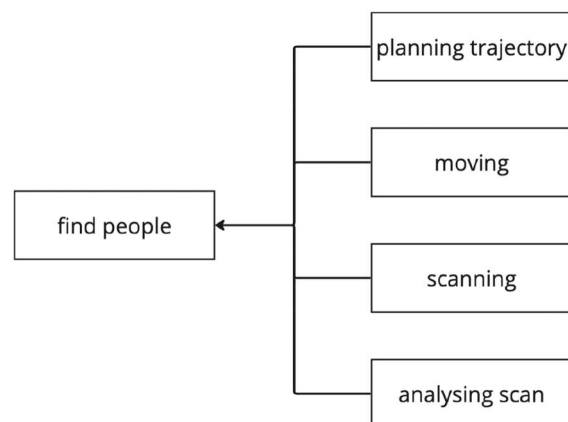
Figure 4. Task of finding people divided into subtasks (which are also atomic tasks).

The first author and a double-coder (non-expert) went through the transcripts and wrote down some codes that came to mind related to comments or questions that may affect the usability of the table. Both coders met to discuss the codes and reach an agreement on the coding scheme. After agreeing on the coding scheme, both coders coded the utterances separately. Both coders met one final time to agree on the coding. During this meeting, some codes were merged.

6. Results

6.1. First phase

In the first phase of the evaluation, most participants showed great interest in using our table for their personal research works. Among other things, participants mentioned our table would be useful in the process of designing their experiments' tasks, calibrating appropriate trust between humans and machines, and designing explanations. We received negative feedback mainly based on the colour code of that version of the table, the inefficiency related to filling in the table, and possible overlapping of dimensions. In particular, it was clear that filling in the first iteration of the table was quite overwhelming for human participants. All feedback from this phase is already integrated in the version of the table presented in this paper. The main change was reducing the size of the table. More concretely, in the first version (in Appendix A.1), we assessed the three dimensions for each possible role (performer with support, independent performer, co-performer, supporter, not involved) for both human and machine, which gives a total of 24 cells to fill in per task. In the second phase, we assess the dimensions only for the possible performers (human independent, machine independent, co-performers) and assume the support feasibility (as explained in Section 3.2.4).

6.2. Thematic analysis (second phase)

The results of the evaluation of our framework are in the feedback given by the participants during the second phase of evaluation. All utterances can be found in our dataset (Centeio Jorge, Jonker, and Tielman 2024a), published in 4TU.ResearchData. We structured this feedback through a thematic analysis, which shows the topics that were brought up throughout the activity and question answering. We calculated the inter-rater reliability for the thematic analysis, resulting in a Cohen's kappa (Landis and Koch 1977) of *0.65* (ran with R package *irr* Gamer, Lemon, and Singh 2012). This value is considered

Table 1. This table shows the number of utterances that were attributed with each of the codes (some utterances were attributed more than one code), and the number of participants that had utterances related to each code. It also shows the total number of attribution of codes in utterances of a certain phase (i.e. during activity, Q1, Q2, Q3 and Q4).

Code ID	Code name	Code count	Phase count	Participants
A1	definition of dimensions	10	50	4
A2	definition of role	23		4
A3	definition of subtasks	11		4
B1	decision-making	4		3
C1	answer granularity	2		2
D1	good structure	4	11	3
D3	clarity	1		1
D4	dimension	2		2
D5	role	1		1
D6	level of detail	2		1
D7	agreement with final results	1		1
E1	unclear definitions	4	9	3
E3	missing evaluation criteria	2		2
E4	context-dependent	3		3
F1	supports planning	5	12	4
F2	discussion starter	3		1
F3	robot design	4		4
G1	rapidly-changing situations	1	4	1
G2	different mindsets	3		2

substantial by Landis and Koch (1977) and *moderate* by McHugh (2012). Because the double-coder was non-expert, and we allowed for more than one code per utterance, this value was considered sufficient to proceed to the analysis. The coding schemes can be found in Figures 5–9, with a respective example of a selected participant's quote (all quotes available in the dataset Centeio Jorge, Jonker, and Tielman 2024a). The respective counts of each code can be found in Table 1.

6.2.1. Activity

During the activity phase, participants were invited to ask questions about the concepts or instructions they could not understand, while being presented the activity. Predominantly, there were questions and discussion regarding the definition of the different **dimensions** (10 utterances), the different **roles** (23 utterances), and the task and its **subtasks** (11 utterances). These continued after the instructions were given, and when the participants were filling the table. All of these codes constitute the theme **Definitions (A0)** (in Figure 5 with respective codes and quotes), which then counts with a total of 34 utterances (as in Table 1). In Table 1, we can also see that each of the codes in theme **A** was attributed to at least four of the five participants.

In particular, the participants had difficulty analysing **dimensions** separately, i.e. not making their analysis of
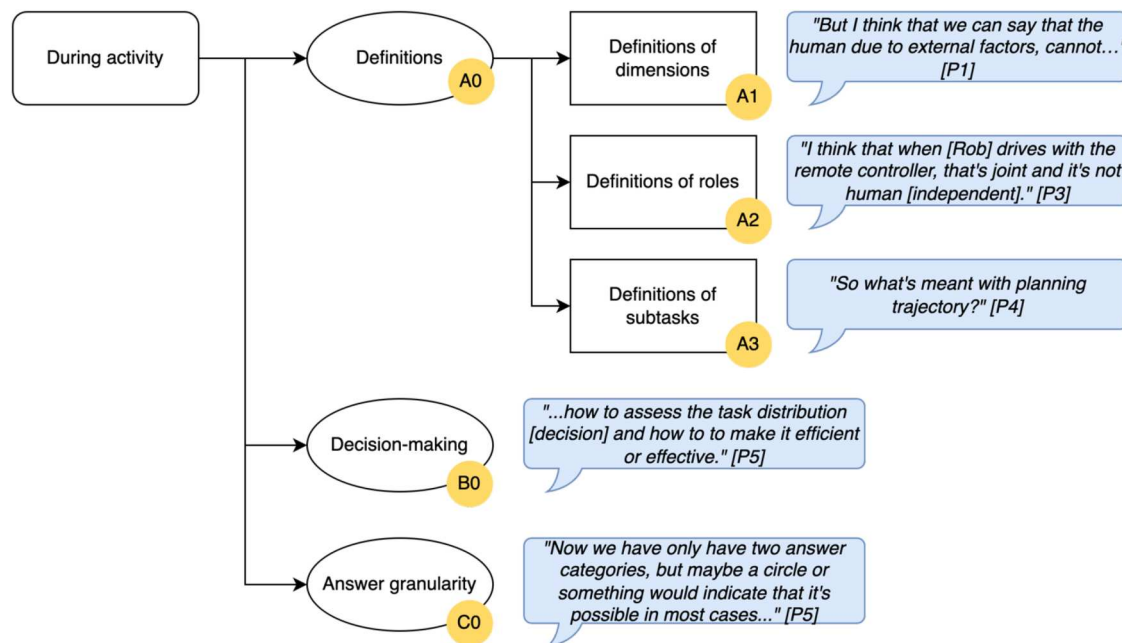
Figure 5. The coding scheme related to the transcripts collected during the activity.

one dimension dependent in another. This can be illustrated by what P4 said,

'I don't know how the "will", the intention, if the external factors weren't there, then he or she would have the intention to do that, but I don't know how to understand the "will have to", whether they could take external factors into account or not.'.

The group also showed difficulty in distinguishing the different **roles**, which can be exemplified by what P3 said, i.e. 'I think that when Rob drives with the remote controller, that's joint [performer] and it's not human [performer]'. Finally, as P1 said 'Yeah, but that's planning trajectory [and not moving], turns out.', the participants showed surprise and difficulty in distinguishing the different **subtasks** and what they involved. Often times, confusion regarding subtask definition led to confusion in roles and even dimensions, which meant that several utterances in this phase were coded with more than one code of theme **definitions (A0)**.

Besides verbalising difficulty with definitions, several participants also gave their opinion on the framework, both while receiving the instructions and filling in the table, sometimes adding suggestions and asking deeper questions about the use of the framework. These were mainly about the decision-making process (**theme Decision-making (B0)**), which reflected two main concerns from three participants: what information distinguishes two or more feasible options (to do a certain task) when someone needs to make a decision

using the table, and how to optimise the decisions made, and how to evaluate the decisions once they are made. Furthermore, two participants suggested that the table could have higher **answer granularity (C0)**, allowing for answers besides yes or no.

6.2.2. Positives (Q1)
In Figure 6, we can see the codes and exemplary quotes of the answers to Q1, when we openly asked participants what they liked the most about our framework. Participants mainly mentioned elements of **table composition (D0)**, which counted with 10 utterances. These included compliments to the **good structure** of the table, the **level of detail**, and its **clarity**. Although in the previous phase, participants showed some difficulty with the definition of the different **dimensions** and **roles**, they mentioned these elements as positives of the framework. One participant also mentioned that they liked that, although there was a lot of discussion, in the end, they all **agreed with the final result** of the table.

6.2.3. Negatives (Q2)
When questioned about the things they did not like (Q2), participants mentioned the **unclear definitions**, which aligns with the results we got in activity phase, where participants discussed the meaning and distinction of roles, dimensions and subtasks. Furthermore, they also recalled the need for **evaluation criteria**, which also reflects the decision-making (B0) in activity. Lastly, participants also showed concern regarding the
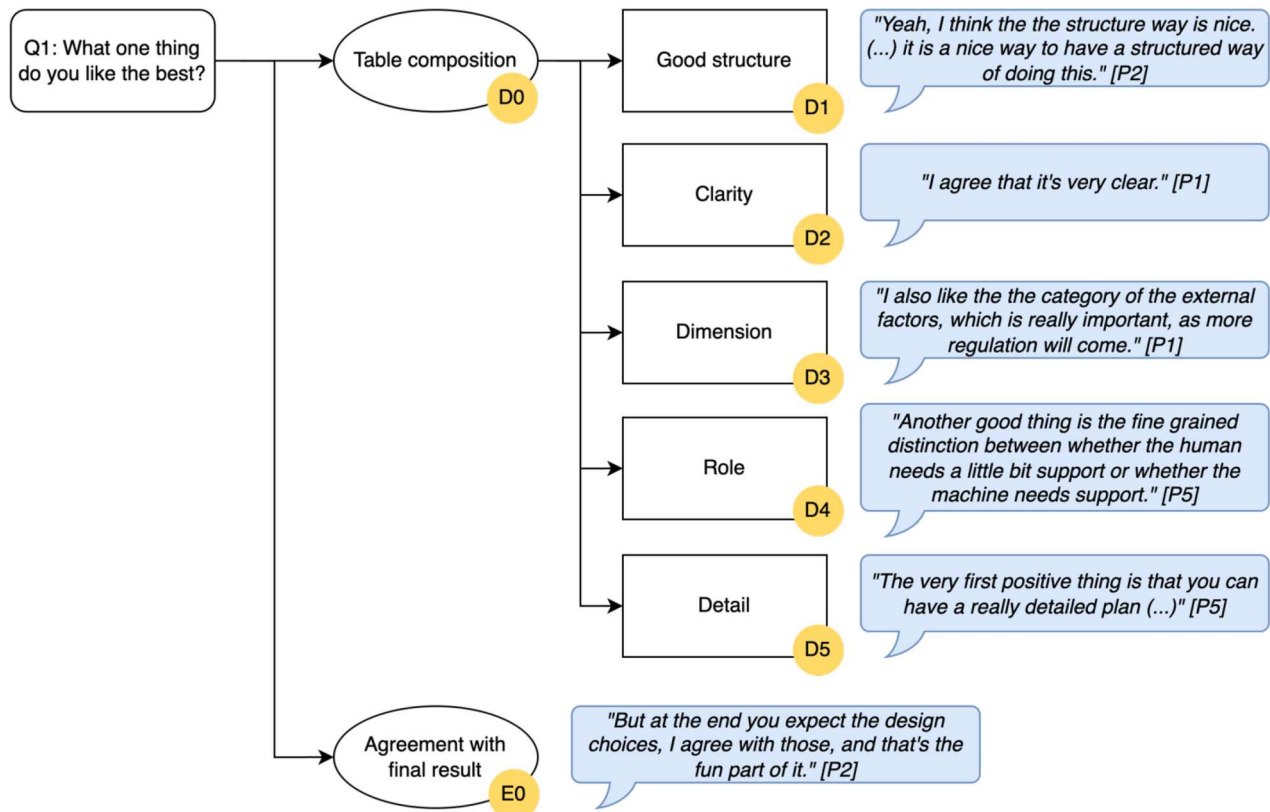
Figure 6. The coding scheme related to the transcripts that answer the question '*What one thing do you like the best?*' (Q1). The blue speech balloons show an utterance that was coded with the corresponding code.

**context dependence** of the table. All codes and exemplary quotes can be found in Figure 7.

6.2.4. When to use (Q3)
Four of the five participants verbalised that our framework **supports [teamwork and task] planning** and that, similarly, it can be used to **design the robot** or AI required for a specific human–machine scenario or task. One participant also mentioned that the use of

the framework is a good **discussion starter**. These codes can be found in Figure 8.

6.2.5. When not to use (Q4)
When asked when they would not use the table, participants were less verbal. However, one participant referred they would not use the table in **rapidly-changing situations** (related to the context-dependency, H0, concerning Q2). Two participants also mentioned
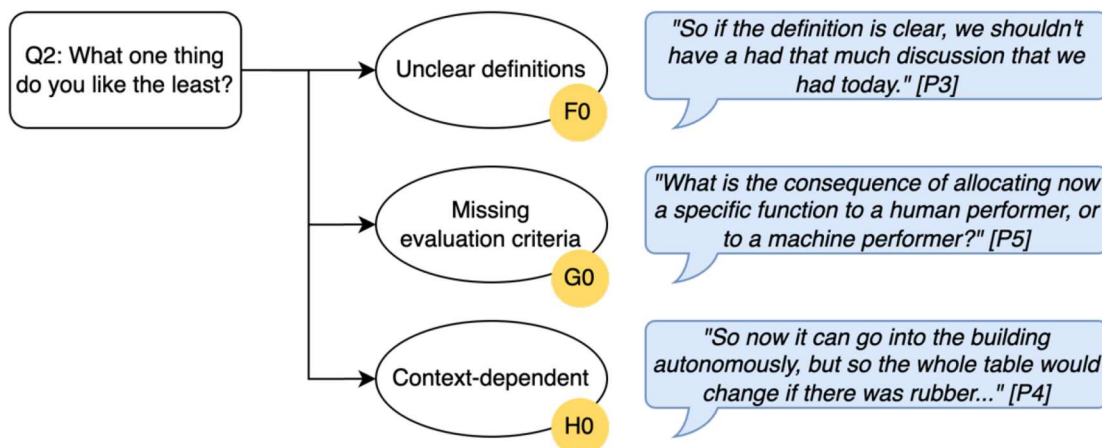
Figure 7. The coding scheme related to the transcripts that answer the question '*What one thing do you like the least?*' (Q2). The blue speech balloons show an utterance that was coded with the corresponding code.
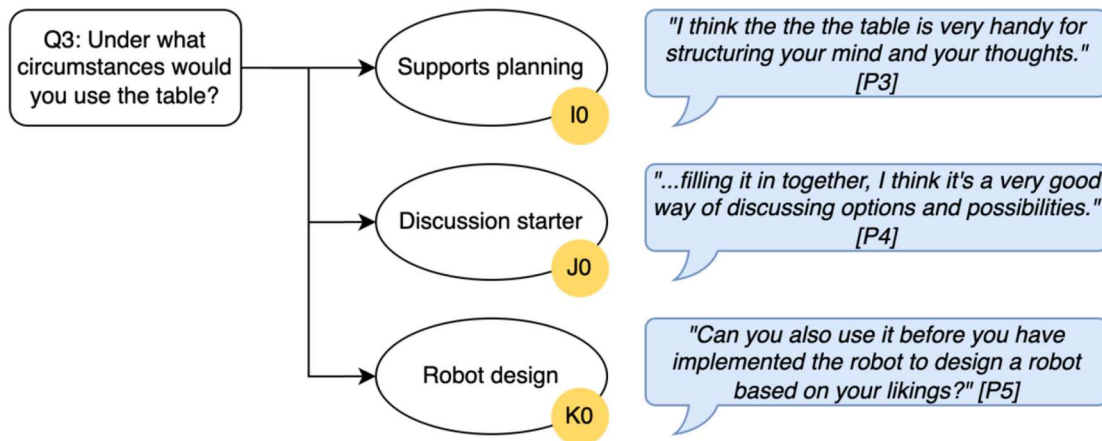
Figure 8. The coding scheme related to the transcripts that answer the question *'Under what circumstances would you use the table?'* (Q3). The blue speech balloons show an utterance that was coded with the corresponding code.

that it might not be feasible to use with people with **different mindsets**, meaning that some workers may not have the capacity to sit down and use such a framework beforehand. Figure 9 shows the exemplary quotes.

6.3. Summary of results

Our results were overall positive, counting with more positive comments than negative in the open answers. The participants were able to use the table as intended and could agree on the final result (code E0). They found it useful for planning (I0), discussing possibilities (J0) and designing artificial teammates (K0). Participants also extensively complimented the table composition (D0), including the chosen dimensions (D3) and possible configurations (D5).

However, there were also some persistent concerns reflected in the participants' comments and questions. They were mainly concerned with (1) the definitions of the dimensions and interdependencies when related to a certain task, (2) the process of filling the table, in particular its (in)efficiency and comprehensive

information and (3) the use of such information not being enough to make decisions. We discuss these results in the next section.

7. Discussion

7.1. Reflection on results and theoretical implications

7.1.1. External factors

Our results build on existing evidence (as in Johnson et al. 2014) that an interdependence analysis can help human-AI team designers identify interdependence relationships in a joint activity. Four out of five participants mentioned that such a tool supports teamwork and coordination of offline planning. Other works have proposed automatic ad-hoc planning for human-AI teamwork, based on teammates' task-based competence (Ali et al. 2022; Azevedo-Sa, Yang, et al. 2021), and role preference (Noormohammadi-Asl et al. 2022, 2023). Our results defend that environment characteristics (external factors dimension) should also be included in task-allocation methods as ethical concerns increase, as
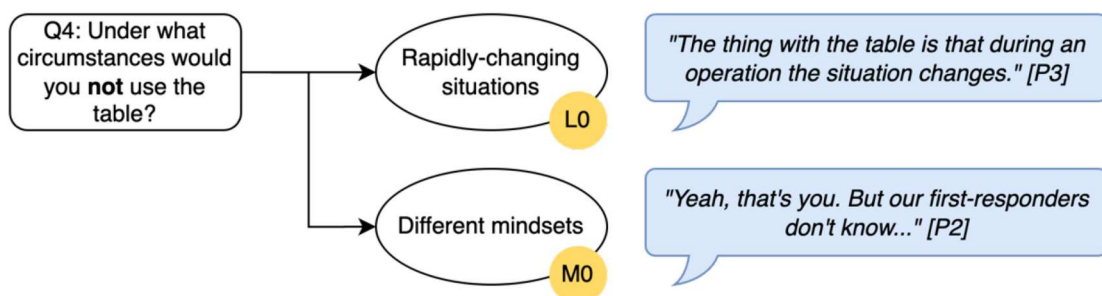
Figure 9. The coding scheme related to the transcripts that answer the question *'Under what circumstances would you not use the table?'* (Q4). The blue speech balloons show an utterance that was coded with the corresponding code.

well as the appearance of new laws related to these. In particular, machines should not make all the decisions. This is supported by two of the participants' interventions, which mentioned that the dimension of the external factors was helpful for task planning. P5 said

> 'The very first positive thing is that you can have a really detailed plan on what sub functions are needed in order to accomplish an overall task and based on capacities or as you called it, external limits or the external environment, you have a good indicator of where you bet your money on.'.

Furthermore, P1 said that

> 'I also like the category of the external factors, which is really important, I think, as more regulation will come also in terms of the communication to, for example, people who want to deploy like, see opportunities in human machine teams, but then due to the AI act, it's not possible any more.',

corroborating that this dimension should be included in human-AI teamwork design and planning. This is aligned with (van der Waa et al. 2021), which presents a dynamic moral task allocation method, and implies that further research on how to integrate ethical and legal boundaries in human–machine teamwork is required.

7.1.2. Communication

As just seen, it can more and more often happen that an AI teammate is capable and willing, but not allowed, as P1 said, 'It's a very nice distinction for communicating that, yes, it's able to, and it's willing to, but we just cannot let that AI do that right now'. This brings our attention to the need to communicate the different dimensions that contribute to a machine not being able to perform a certain task. Although communication does not appear explicitly in our codes, it does come implicit in some of them, all of them mentioned as positive characteristics, such as clarity (D2), agreement with final result (E0) and discussion starter (J0). Although there has been an effort to explain the automation's mental states to the human during and after collaboration (see e.g. Le Guillou, Prévot, and Berberian 2023; Luebbers et al. 2023; Tabrez 2024), bidirectional communication should be explicitly included in teamwork design and task planning methods. Further research is required to investigate how this can be done naturally between the human and the machine.

7.1.3. Definitions

Both during the presentation of the table and the activity, participants asked about the definitions of dimensions, roles and subtasks (A0-A3 and F0). In

particular, the most predominant concern was related to the definition of a certain role for a certain task. For example, P1 said, 'Teleoperating sounds like human support and not joint.' We recognise a difficulty in defining what support and co-perform means, depending on the task. In the case of the task movement, the robot can be teleoperated or move autonomously. If a human teleoperates the robot, does it mean the human and the machine do it jointly (they're co-performers)? Or is the human the only one performing this action? Or one is performing and the other supporting, and if so who is what? We believe this difficulty comes mainly from a lack of precision on what each task means. In these cases, users should try to divide the tasks into even smaller subtasks so that the roles become clearer. For example, perhaps if we were to have an atomic task deciding next movement and another physically changing positions, it would be clearer that both the robot and the human can decide the next movement (independently) but only the robot can actually change its own physical position because the human is not physically there to do it. Our findings suggest that existing human–machine models need to be updated in order to incorporate team configurations that are not binary (e.g. leader vs followed as in Noormohammadi-Asl et al. 2022). These human–machine models should also account for the fact that team members having different natures alters the interpretation and meaning of the task (e.g. works that assume the same task definition for humans and machines, such as Ali et al. 2022; Azevedo-Sa, Yang, et al. 2021; Johnson et al. 2014).

7.1.4. Filling in the table

In the famous Technology Acceptance Model (TAM) (Davis 1989), Davis proposes that perceived usefulness and perceived ease-of-use are the two main factors that influence the actual use of a technology. The author defines perceived usefulness as 'the degree to which a person believes that using a particular system would enhance his or her job performance' and perceived ease-of-use as 'the degree to which a person believes that using a particular system would be free of effort' (p. 320). Our results show that it is not always easy to increase usefulness without increasing effort, and vice versa. In our case, asking the user for more information to include in the table increases usefulness (as more information can lead to better decisions), however, more information means that the user needs to fill in a higher number of cells in the table, which leads to a higher effort to the user, which decreases the perceived ease-of-use.

In order to increase the perceived ease-of-use, we decided to make the cells binary (check or cross), mostly because we believed this would be easier for a person than to come up with a value from a certain range (let's say from 0 to 5). Interestingly, this was pointed out by a participant (C0), who felt the need to express something other than the possible answers (check or cross). However, of course a binary value gives us way less information than a wider range (which may decrease the perceived usefulness). Similarly, although participants were generally happy with the level of detail (D5), it was also brought up, as a negative aspect of the table, that the table is context-dependent (H0), i.e. that changing the context means changing the values in the table. We understand how a user can perceive this as a negative point, since they would have to fill in the table again if the context changes, decreasing their perceived ease-of-use. However, it is hard to make the table context-free (which improves the ease-of-use) while not decreasing its perceived usefulness, i.e. having enough information about team members' competence, willingness and external factors, in a specific context. For example, a machine may be allowed to hold a knife if a human is not present, but not otherwise. This means that the value in the external factors dimension is going to change depending on the context, requiring more information. Nevertheless, we need that information to know whether we can give the task of chopping potatoes to the machine.

Actually, the fact that the table is context-dependent increased the perceived usefulness according to other participants (D3), for example, we have mentioned before that some participants highly appreciated the external factors dimension of the table. The external factors dimension is the most context-dependent dimension of the table. Trying to accommodate the HO, we believe it is possible to reduce the effort from the users in real-time, so that the users do not to change the values in the table whenever the context changes. This can be done by discriminating beforehand all possible contexts and including this context in the atomic tasks. For example, we could have the task *holding a knife when a human is around*, which will not change depending on the context (since it is in the description of the task). However, we still need a way of deciding which atomic tasks are used, which still depends on the context.

With the two examples given in this subsection, i.e. the level of detail in the cells of the table and the context-dependence, we realise that some participants may feel like they need to put in a lot of work before the table becomes useful. Furthermore, as system designers, we also need to consider that requiring loads of information from the user does not only harm the user's perceived effort, but it may also decrease the overall efficiency of the system (since it may take a lot of time to disclose this information, for example, which may not be possible in real-time). This poses a challenge to AI designers that need to ensure the compliance with new regulations of transparency, traceability and accountability that are emerging around the world, e.g. the European AI Act,[1] and the IEEE Ethically Aligned Design.[2] If we want to have a reliable and transparent framework, that acknowledges its context and adapts to the circumstances, we may have to disclose a high load of contexts and nuances, which require a high effort from the user. This information is crucial to properly explain and justify decisions made by agents that possibly use the framework, such as explaining that they cannot help with chopping because a human is around, and that in such cases they are not allowed to hold a knife. We expect these challenges to be more and more present in the development of human–machine systems and collaboration design, as regulations become stricter. These findings show that there is a need for researching guided conversation to fill in these tables (also Johnson and Bradshaw 2021; Johnson et al. 2014), making the process of human–machine team design effortless to the human, while guaranteeing ethical compliance.

7.1.5. Making decisions

The final topic of concern had to do with decision-making itself. Although participants saw value in the table to help to make decisions (I0, J0), it was mentioned that an evaluation criteria was missing. Participants felt the need to have further information about what was the goal of the task allocation, as well as what each subtask meant for the achievement of that goal. For example, P5 asked '*What is the consequence of allocating now a specific function to a human performer, or to a machine performer?*' Indeed, there may be different objectives when allocating tasks in human–machine teams, including reduction of the non-ergonomic human task, productivity and human satisfaction (Nikolakis et al. 2018). We believe this information is important, but we decided to keep it out of this version of the table. The main reason is that the information that is necessary to make decisions, such as expected workload for each teammate, or total time per team configuration, is hard to predict and obtain. In fact, for the first phase of evaluation, we prepared a mock side table with this type of information, to be used before the decision-making step. We learnt from the participants that this would be very hard to actually obtain for real-life scenarios, e.g. how to calculate the workload of a human chopping ten potatoes? As such, this poses entirely different questions

than the ones we are trying to answer in this paper. These findings suggest that existing task allocation and decision models (such as Ali et al. 2022; Bhat et al. 2022; Unhelkar, Li, and Shah 2020) should include human-centred factors for optimisation and utility calculation (reward and penalty), such as accounting for the system values (see e.g. Harbers and Neerincx 2017) and major risks, which should be changeable from context to context.

7.1.6. Summary of theoretical implications

Results show that experts value our framework and believe it supports human–machine teamwork planning, discussion and design. The table is successful in improving communication and supports the team design with machines (and artificial intelligence) with variable levels of autonomy and permission (which is independent of the machine's capabilities). This suggests that Interdependence and Trust Analysis could benefit planning and designing of human–machine teams for different contexts such as disaster response (De Greeff et al. 2018), search and rescue (Saad, Hindriks, and Neerincx 2018), cooking (Goubard and Demiris 2023), driving (Azevedo-Sa, Jayaraman, et al. 2021) and healthcare (Cypko et al. 2022). However, we found that the interpretation of the table's roles and subtasks can be challenging and subjective, which suggests that tasks' and roles' definitions should depend on the natures of the agents involved (e.g. different embodiment and cognitive characteristics lead to different meanings of *support*). We also see a trade-off between efficiency (effortlessness) and completeness (usefulness), as participants appreciate the level of detail and information of the table, but are not very happy about having to provide so much information during the analysis. This suggests a need for more natural communication to fill in the table (instead of going through the whole table cell by cell), such as guided dialogues. Finally, the focus group showed a need to include different optimisation human-centred policies, depending on the context, in existing decision-making models for human–machine collaboration. For example, in some scenarios it might be important to reward a certain value (e.g. safety) more than others (e.g. privacy), or simply maximise for user's satisfaction.

7.2. Limitations and future work

Our work and method present some limitations that also open ways for improvement in future work. As we mentioned earlier in this section, there is a trade-off between efficiency (effortlessness) and completeness (usefulness), which may impact the reliability and adaptability of the system. This being said, we had to compromise on the amount of information included in the table. Such decisions also led to assumptions which may be seen as limitations. In particular, we had to reduce the rows of the table, which led to assuming that the feasibility of an agent's support could be inferred from the feasibility of that agent's independence. Ideally, we would have a separate analysis for support, but that proved to be overwhelming to participants. However, there may be applications in which an extended version of the table (as in Appendix A.1) can be more suitable, and so users can still use a broader version of the table. In fact, in future work, we would like to implement artificial agents that use the table as a support for task selection and allocation (stage 2 in Section 4). In such cases, the agent needs to form and update beliefs about all dimensions, all teammates, all tasks and respective interdependence roles. It is surely less overwhelming for an artificial agent than for a human to deal with a bigger table, while at the same time more necessary, as that table will explicitly include the important information. After stage 2, we also want to explore learning algorithms that update the table automatically throughout interactions. Furthermore, we would also like to implement an automatic generation of explanations and/or justifications for the AI that makes use of this framework for task allocation or selection.

Another possible limitation is that this interdependence and trust analysis (ITA) table assumes that an agent that uses it has full knowledge, i.e. enough information regarding all agents and all dimensions, to fill in the table. We have not accounted for cases in which there is no such knowledge, and what that means in terms of feasibility. In future work, we would like to accommodate this option. It would also be relevant to find a way to represent the accuracy of each cell. For example, perhaps an agent believes that the other can do a task independently, but is not 100% sure. This may affect their future decisions, so it should be represented, as it will affect the future risk of the decision. Overall, risk is not included in this analysis. Besides accuracy, we can also see the risk of going for a specific design choice, and even the risk of *not* going for a specific design choice. It has been mentioned that one of the participants' concerns was how to make a decision after knowing which configurations are feasible. We have mentioned that there might be several criteria that would prioritise some choices over others, and risk is definitely one of them. However, risk is also hard to calculate and assumes there is knowledge for that, so for simplification, we did not include it.

Finally, the thematic analysis used for evaluating the table has its limitations, such as possible bias and

dialogue manipulation of a more leading or dominating participant (Gundumogula and Gundumogula 2020). Although we made sure to ask the questions to each participant, one's answers are naturally affected by the others'. The focus group of the second phase of the analysis was composed of five people, which is considered by some authors to be enough (Gundumogula and Gundumogula 2020), but some others consider it to be too small of a group (Morgan 1997). We acknowledge the limitation of the small sample size of the two dyadic interviews and the focus group. Although this method gives us an initial understanding of how experts perceive our framework, further research is necessary to study the extent to which these insights transfer to other groups of human–machine team designers. The thematic analysis inter reliability was considered sufficient, as the double coder was not an expert in human–machine teamwork and utterances allowed multiple codes. Most disagreements were in the cases of multiple codes, especially in the ones that included definitions of dimensions, roles or subtasks (theme *definitions*). For example, in the quote 'So in case of or for the subtask peeling, the human doesn't want to…but how is it then a mandatory joint?', we see dimensions, tasks and roles being mentioned. It can be hard to decide what is the most important code(s) for such utterance. In future work, we want to run a more objective evaluation of the table in a more involving scenario.

8. Conclusion

In this paper, we present an extension of the Interdependence Analysis for human–machine teams. Our approach includes a discriminated analysis of the trustworthiness dimensions of competence (i.e. skills, knowledge), willingness (i.e. intention, preference) and external factors (i.e. opportunity, resources), for each possible team interdependence configuration, for each subtask. This table can support the design of human–machine teams, including the allocation of tasks. In fact, it can also be used for decision-making of team members, either human or machine, supporting task selection too. By using this table as a shared mental model, decisions may become more transparent, justifiable and interpretable, which may lead to an increased and appropriate trust among teammates.

Notes

1. https://artificialintelligenceact.eu/.
2. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.

References

Abuhaimed, Sami, Selim Karaoglu, and Sandip Sen. 2023. "Choosing the Task Allocator: Effect on Performance and Satisfaction in Human-Agent Team." In *The International FLAIRS Conference Proceedings*, Vol. 36.

Ali, Arsha, Hebert Azevedo-Sa, Dawn M. Tilbury, and Lionel P. Robert Jr. 2022. "Heterogeneous Human–Robot Task Allocation Based on Artificial Trust." *Scientific Reports* 12 (1): 15304. https://doi.org/10.1038/s41598-022-19140-5.

Azevedo-Sa, Hebert, Suresh Kumaar Jayaraman, Connor T. Esterwood, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. 2021. "Real-Time Estimation of Drivers' Trust in Automated Driving Systems." *International Journal of Social Robotics* 13 (8): 1911–1927. https://doi.org/10.1007/s12369-020-00694-1.

Azevedo-Sa, Hebert, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. 2021. "A Unified Bi-Directional Model for Natural and Artificial Trust in Human–Robot Collaboration." *IEEE Robotics and Automation Letters* 6 (3): 5913–5920. https://doi.org/10.1109/LRA.2021.3088082.

Baum, Kevin, Joanna Bryson, Frank Dignum, Virginia Dignum, Marko Grobelnik, Holger Hoos, Morten Irgens, et al. 2023. "From Fear to Action: AI Governance and Opportunities for All." *Frontiers in Computer Science* 5:1210421. https://doi.org/10.3389/fcomp.2023.1210421.

Bhat, Shreyas, Joseph B. Lyons, Cong Shi, and X. Jessie Yang. 2022. "Clustering Trust Dynamics in a Human–Robot Sequential Decision-Making Task." *IEEE Robotics and*

Automation Letters 7 (4): 8815–8822. https://doi.org/10.1109/LRA.2022.3188902.

Bradshaw, Jeffrey M., Paul J. Feltovich, Hyuckchul Jung, Shriniwas Kulkarni, William Taysom, and Andrzej Uszok.2004. "Dimensions of Adjustable Autonomy and Mixed-Initiative Interaction, edited by *Agents and Computational Autonomy. AUTONOMY 2003. Lecture Notes in Computer Science*. Vol. 2969, edited by M. Nickles, M. Rovatsos, and G. Weiss. , Melbourne, VIC, Australia: Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-25928-2_3.

Braun, Virginia, and Victoria Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3 (2): 77–101. https://doi.org/10.1191/1478088706qp063oa.

Carroll, Micah, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca D. Dragan. 2019. "On the Utility of Learning about Humans for Human-AI Coordination." In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*, edited by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, 5175–5186. https://proceedings.neurips.cc/paper/2019/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html.

Castelfranchi, Cristiano, and Rino Falcone. 2010. "Trust & Self-Organising Socio-technical Systems." In *Trustworthy Open Self-Organising Systems*, edited by Wolfgang Reif, Gerrit Anders, Hella Seebach, Jan-Philipp Steghöfer, Elisabeth André, Jörg Hähner, Christian Müller-Schloer, and Theo Ungerer, 209–229. Birkhäuser, Cham: Autonomic Systems. https://doi.org/10.1007/978-3-319-29201-4_8.

Centeio Jorge, Carolina, Nikki H. Bouman, Catholijn M. Jonker, and Myrthe L. Tielman. 2023. "Exploring the Effect of Automation Failure on the Human's Trustworthiness in Human-Agent Teamwork." *Frontiers in Robotics and AI* 10:1143723. https://doi.org/10.3389/frobt.2023.1143723.

Centeio Jorge, Carolina, Catholijn M. Jonker, and Myrthe L. Tielman. 2023. "Artificial Trust for Decision-Making in Human-AI Teamwork: Steps and Challenges." In *CEUR Workshop Proceedings*, Vol. 3456.

Centeio Jorge, Carolina, C. M. Jonker, and Myrthe Tielman. 2024a. "Data for Interdependence and Trust Analysis (ITA): A Framework for Human–Machine Team Design". https://doi.org/10.4121/998296ec-7696-4180-8d7e-9af8588b1182.v1.

Centeio Jorge, Carolina, Catholijn M. Jonker, and Myrthe L. Tielman. 2024b. "How Should An AI Trust Its Human Teammates? Exploring Possible Cues of Artificial Trust." *ACM Transactions on Interactive Intelligent Systems* 14 (1): 1–26. https://doi.org/10.1145/3635475.

Cervantes, José-Antonio, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos. 2020. "Artificial Moral Agents: A Survey of the Current Status." *Science and Engineering Ethics* 26 (2): 501–532. https://doi.org/10.1007/s11948-019-00151-x.

Cypko, Mario A., Lea Timmermann, Igor M. Sauer, and Claudia Müller-Birn. 2022. "Towards Human–Robotic Collaboration: Observing Teamwork of Experienced Surgeons in Robotic-Assisted Surgery." In *MuC '22: Mensch und Computer 2022, Darmstadt Germany, September 4–7, 2022*, edited by Max Mühlhäuser, Christian Reuter, Bastian Pfleging, Thomas Kosch, Andrii Matviienko, Kathrin Gerling, Sven Mayer, Wilko Heuten, Tanja Döring, Florian Müller, and Martin Schmitz, 566–571. ACM. https://doi.org/10.1145/3543758.3549891.

Davis, Fred D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." *MIS Quarterly* 13 (3): 319–340. https://doi.org/10.2307/249008.

De Greeff, Joachim, Tina Mioch, Willeke Van Vught, Koen Hindriks, Mark A. Neerincx, and Ivana Kruijff-Korbayová. 2018. "Persistent Robot-Assisted Disaster Response." In *Companion of the 2018 ACM/IEEE International Conference on Human–Robot Interaction*, 99–100.

De Visser, Ewart J., Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. "Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams." *International Journal of Social Robotics* 12 (2): 459–478. https://doi.org/10.1007/s12369-019-00596-x.

Fahnenstich, Hannah, Tobias Rieger, and Eileen Roesler. 2023. "Trusting under Risk–comparing Human to AI Decision Support Agents." *Computers in Human Behavior* 153 (C). https://doi.org/10.1016/j.chb.2023.108107..

Falcone, Rino, and Cristiano Castelfranchi. 2004. "Trust Dynamics: How Trust Is Influenced by Direct Experiences and by Trust Itself." In *3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004), 19–23 August 2004, New York, NY, USA*, 740–747. IEEE Computer Society.

Falcone, Rino, Michele Piunti, Matteo Venanzi, and Cristiano Castelfranchi. 2013. "From Manifesta to Krypta: The Relevance of Categories for Trusting Others." *ACM Transactions on Intelligent Systems and Technology* 4 (2): 27:1–27:24. https://doi.org/10.1145/2438653.2438662.

Gamer, Matthias, Jim Lemon, and Ian Fellows Puspendra Singh. 2012. *irr: Various Coefficients of Interrater Reliability and Agreement*. https://CRAN.R-project.org/package=irr. R package version 0.84.1.

Georgeff, Michael P., Barney Pell, Martha E. Pollack, Milind Tambe, and Michael J. Wooldridge. 1998. "The Belief-Desire-Intention Model of Agency." In *Intelligent Agents V, Agent Theories, Architectures, and Languages, 5th International Workshop, ATAL '98, Paris, France, July 4–7, 1998, Proceedings (Lecture Notes in Computer Science, Vol. 1555)*, edited by Jörg P. Müller, Munindar P. Singh, and Anand S. Rao, 1–10. Springer. https://doi.org/10.1007/3-540-49057-4_1.

Gervits, Felix, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. 2020. "Toward Genuine Robot Teammates: Improving Human–Robot Team Performance Using Robot Shared Mental Models." In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 429–437.

Goubard, Cedric, and Yiannis Demiris. 2023. "Cooking Up Trust: Eye Gaze and Posture for Trust-Aware Action Selection in Human–Robot Collaboration." In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems (Edinburgh, United Kingdom,) (TAS '23)*, Article 34, 5. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3597512.3597518.

Griffiths, N. 2005. "Task Delegation Using Experience-Based Multi-Dimensional Trust." In *AAMAS '05*.

Gundumogula, Manju, and M. Gundumogula. 2020. "Importance of Focus Groups in Qualitative Research." *International Journal of Humanities and Social Science (IJHSS)* 8 (11): 299–302.

Hancock, Peter A., Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart de Visser, and Raja Parasuraman.. 2011. "A Meta-Analysis of Factors Affecting Trust in Human–Robot Interaction." *Human Factors* 53 (5): 517–527. https://doi.org/10.1177/0018720811417254.

Harbers, Maaike, and Mark A. Neerincx. 2017. "Value Sensitive Design of a Virtual Assistant for Workload Harmonization in Teams." *Cognition, Technology & Work* 19 (2–3): 329–343. https://doi.org/10.1007/s10111-017-0408-4.

Hoesterey, Steffen, and Linda Onnasch. 2023. "The Effect of Risk on Trust Attitude and Trust Behavior in Interaction with Information and Decision Automation." *Cognition, Technology & Work* 25 (1): 15–29. https://doi.org/10.1007/s10111-022-00718-y.

Johnson, Matthew, and Jeffrey M. Bradshaw. 2021. "Chapter 16 - The Role of Interdependence in Trust." In *Trust in Human-Robot Interaction*, edited by Chang S. Nam and Joseph B. Lyons, 379–403. Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00016-2.

Johnson, Matthew, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna Van Riemsdijk, and Maarten Sierhuis. 2014. "Coactive Design: Designing Support for Interdependence in Joint Activity." *Journal of Human–Robot Interaction* 3 (1): 43–69. https://doi.org/10.5898/JHRI.3.1.Johnson.

Klien, Glen, David D. Woods, Jeffrey M. Bradshaw, Robert R. Hoffman, and Paul J. Feltovich. 2004. "Ten Challenges for Making Automation a "team Player" in Joint Human-Agent Activity." *IEEE Intelligent Systems* 19 (6): 91–95. https://doi.org/10.1109/MIS.2004.74.

Krueger, Richard A., Mary Anne Casey, Jonathan Donner, Stuart Kirsch, and Jonathan N. Maack. 2002. *Social Analysis: Selected Tools and Techniques (English)*. Social Development Papers; Vol. 36. Washington, D.C: World Bank Group. http://documents.worldbank.org/curated/en/568611468763498929/Social-analysis-selected-tools-and-techniques.

Landis, J. Richard, and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–174(16 pages). https://doi.org/10.2307/2529310.

Law, Theresa, and Matthias Scheutz. 2021. "Chapter 2 - Trust: Recent Concepts and Evaluations in Human-Robot Interaction." In *Trust in Human-Robot Interaction,*, edited by Chang S. Nam and Joseph B. Lyons, 27–57. Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00002-2.

Lee, J. D., and Katrina A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors: The Journal of Human Factors and Ergonomics Society* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50.30392.

Lee, Sun Kyong, and Juhyung Sun. 2023. "Testing a Theoretical Model of Trust in Human–machine Communication: Emotional Experience and Social Presence." *Behaviour & Information Technology* 42 (16): 2754–2767. https://doi.org/10.1080/0144929X.2022.2145998.

Le Guillou, Marin, Laurent Prévot, and Bruno Berberian. 2023. "Trusting Artificial Agents: Communication Trumps Performance." In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023–2 June 2023*, edited by Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh, 299–306. ACM. https://doi.org/10.5555/3545946.3598651.

Lewis, Michael, Huao Li, and Katia Sycara. 2021. "Chapter 14 - Deep learning, Transparency, and Trust in Human Robot Teamwork." In *Trust in Human-Robot Interaction*, edited by Chang S. Nam and Joseph B. Lyons, 321–352. Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00014-9.

Luebbers, Matthew B., Aaquib Tabrez, Kyler Ruvane, and Bradley Hayes. 2023. "Autonomous Justification for Enabling Explainable Decision Support in Human–Robot Teaming." In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10–14, 2023*, edited by Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu. https://doi.org/10.15607/RSS.2023.XIX.002.

Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. "An Integrative Model of Organizational Trust." *Source: The Academy of Management Review* 20 (3): 709–734.

McHugh, Mary L. 2012. "Interrater Reliability: The Kappa Statistic." *Biochemia Medica* 22 (3): 276–282. https://doi.org/10.11613/issn.1846-7482.

McKee, Kevin R., Xuechunzi Bai, and Susan T. Fiske. 2022. "Warmth and Competence in Human-Agent Cooperation." In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (Virtual Event, New Zealand) (AAMAS '22)*, 898–907. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Mehrotra, Siddharth, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. "Integrity-Based Explanations for Fostering Appropriate Trust in AI Agents." *ACM Transactions on Interactive Intelligent Systems* 14 (1): 1–36. https://doi.org/10.1145/3610578.

Michie, Susan, Maartje M. Van Stralen, and Robert West. 2011. "The Behaviour Change Wheel: A New Method for Characterising and Designing Behaviour Change Interventions." *Implementation Science* 6 (1): 1–12. https://doi.org/10.1186/1748-5908-6-1.

Morgan, David L. 1997. *Focus Groups as Qualitative Research*. 2nd ed. Thousand Oaks, CA:: SAGE Publications, Inc. https://doi.org/10.4135/9781412984287.

Morgan, David L., Jutta Ataie, Paula Carder, and Kim Hoffman. 2013. "Introducing Dyadic Interviews as a Method for Collecting Qualitative Data." *Qualitative Health Research* 23 (9): 1276–1284. https://doi.org/10.1177/1049732313501889.

Nenna, Federica, Davide Zanardi, Egle Maria Orlando, Michele Mingardi, Giulia Buodo, and Luciano Gamberini. 2024. "Addressing Trust and Negative Attitudes Toward Robots in Human–Robot Collaborative Scenarios: Insights from the Industrial Work Setting." In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2024, Crete, Greece, June 26–28, 2024*, ACM. https://doi.org/10.1145/3652037.3663905.

Nikolakis, Nikolaos, Kostantinos Sipsas, Panagiota Tsarouchi, and Sotirios Makris. 2018. "On a Shared Human–robot

Task Scheduling and Online Re-Scheduling." *Procedia CIRP* 78:237–242. https://doi.org/10.1016/j.procir.2018.09.055.

Noormohammadi-Asl, Ali, Ali Ayub, Stephen L. Smith, and Kerstin Dautenhahn. 2022. "Task Selection and Planning in Human–Robot Collaborative Processes: To Be a Leader or a Follower?" In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1244–1251. IEEE.

Noormohammadi-Asl, Ali, Ali Ayub, Stephen L. Smith, and Kerstin Dautenhahn. 2023. "Adapting to Human Preferences to Lead or Follow in Human–Robot Collaboration: A System Evaluation." In *32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2023, Busan, Republic of Korea, August 28–31, 2023*, 1851–1858. IEEE. https://doi.org/10.1109/RO-MAN57019.2023.10309328.

Rezaei Khavas, Zahra, Monish Reddy Kotturu, S Reza Ahmadzadeh, and Paul Robinette. 2024. "Do Humans Trust Robots that Violate Moral Trust?." *ACM Transactions on Human–Robot Interaction* 13 (2): 1–30. https://doi.org/10.1145/3651992.

Saad, Elie, Koen V. Hindriks, and Mark A. Neerincx. 2018. "Ontology Design for Task Allocation and Management in Urban Search and Rescue Missions." In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence, ICAART 2018, Volume 2, Funchal, Madeira, Portugal, January 16–18, 2018*, edited by Ana Paula Rocha and H. Jaap van den Herik, 622–629. SciTePress. https://doi.org/10.5220/0006661106220629.

Salas, Eduardo, Dana E. Sims, and C. Shawn Burke. 2005. "Is there a "big Five" in Teamwork?." *Small Group Research* 36 (5): 555–599. https://doi.org/10.1177/1046496405277134.

Santoni de Sio, Filippo, and Jeroen van den Hoven. 2018. "Meaningful Human Control over Autonomous Systems: A Philosophical Account." *Frontiers in Robotics and AI* 5:15. https://doi.org/10.3389/frobt.2018.00015.

Seeber, Isabella, Eva Bittner, Robert O. Briggs, Triparna De Vreede, Gert-Jan De Vreede, Aaron Elkins, Ronald Maier, et al. 2020. "Machines as Teammates: A Research Agenda on AI in Team Collaboration." *Information & Management* 57 (2): 103174. https://doi.org/10.1016/j.im.2019.103174.

Sierhuis, Maarten, Jeffrey Bradshaw, Alessandro Acquisti, Ron van Hoof, Renia Jeffers, and Andrzej Uszok. 2003. "Human-Agent Teamwork and Adjustable Autonomy in Practice." In *Proceeding of the 7th International Symposium on Artificial Intelligence, Robotics and Automation in Space*, 1–8.

Song, Yao, and Yan Luximon. 2024. "When Trustworthiness Meets Face: Facial Design for Social Robots." *Sensors* 24 (13): 4215. https://doi.org/10.3390/s24134215.

Stuck, Rachel E., Brittany E. Holthausen, and Bruce N. Walker. 2021. "The Role of Risk in Human–Robot Trust." In *Trust in Human–Robot Interaction*, 179–194. Elsevier.

Stuck, Rachel E., Brianna J. Tomlinson, and Bruce N. Walker. 2022. "The Importance of Incorporating Risk into Human-Automation Trust." *Theoretical Issues in Ergonomics Science* 23 (4): 500–516. https://doi.org/10.1080/1463922X.2021.1975170.

Szulc, Joanna, and Nigel King. 2022. "The Practice of Dyadic Interviewing: Strengths, Limitations and Key Decisions." *Forum Qualitative Sozialforschung Forum: Qualitative Social Research* 23 (2). https://doi.org/10.17169/fqs-22.2.3776.

Tabrez, Aaquib. 2024. "Autonomous Policy Explanations for Effective Human–Machine Teaming." In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20–27, 2024, Vancouver, Canada* edited by Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, 23423–23424. AAAI Press. https://doi.org/10.1609/AAAI.V38I21.30412.

Ulfert, Anna-Sophie, Eleni Georganta, Carolina Centeio Jorge, Siddharth Mehrotra, and Myrthe Tielman. 2023. "Shaping a Multidisciplinary Understanding of Team Trust in Human-AI Teams: A Theoretical Framework." *European Journal of Work and Organizational Psychology* 33 (2): 158–171. https://doi.org/10.1080/1359432X.2023.2200172.

Ullman, Daniel, and Bertram F. Malle. 2020. "Measuring Gains and Losses in Human–Robot Trust: Evidence for Differentiable Components of Trust." In *Proceedings of the 14th ACM/IEEE International Conference on Human–Robot Interaction (Daegu, Republic of Korea) (HRI '19)*, 618–619. IEEE Press.

Unhelkar, Vaibhav V., Shen Li, and Julie A. Shah. 2020. "Decision-Making for Bidirectional Communication in Sequential Human–Robot Collaborative Tasks." In *HRI '20: ACM/IEEE International Conference on Human–Robot Interaction, Cambridge, United Kingdom, March 23–26, 2020*, edited by Tony Belpaeme, James E. Young, Hatice Gunes, and Laurel D. Riek, 329–341. ACM. https://doi.org/10.1145/3319502.3374779.

van de Kieft, Iris, Catholijn M. Jonker, and M. Birna van Riemsdijk. 2011. "Shared Mental Models for Decision Support Systems and Their Users." In *3rd International Workshop on Collaborative Agents-REsearch and development (CARE 2011)*, 54–63.

van den Bosch, Karel, Tjeerd Schoonderwoerd, Romy Blankendaal, and Mark Neerincx. 2019. "Six Challenges for Human-AI Co-Learning." In *Adaptive Instructional Systems: First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*, 572–589. Springer.

van der Waa, Jasper, Sabine Verdult, Karel van den Bosch, Jurriaan van Diggelen, Tjalling Haije, Birgit van der Stigchel, and Ioana Cocu. 2021. "Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations." *Frontiers in Robotics and AI* 8:640647. https://doi.org/10.3389/frobt.2021.640647.

van Diggelen, Jurriaan, and Matthew Johnson. 2019. "Team Design Patterns." In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 118–126.

Van Zoelen, Emma, Tina Mioch, Mani Tajaddini, Christian Fleiner, Stefani Tsaneva, Pietro Camin, Thiago S. Gouvêa, Kim Baraka, Maaike H. T. De Boer, and Mark A. Neerincx. 2023. "Developing Team Design Patterns for Hybrid Intelligence Systems." In *HHAI 2023: Augmenting Human Intellect - Proceedings of the 2nd International Conference on Hybrid Human-Artificial Intelligence.*

Frontiers in Artificial Intelligence and Applications; Vol. 368, 3–16. IOS Press. https://doi.org/10.3233/FAIA230071.

Verhagen, Ruben S., Siddharth Mehrotra, Mark A. Neerincx, Catholijn M. Jonker, and Myrthe L. Tielman. 2022. Exploring Effectiveness of Explanations for Appropriate Trust: Lessons from Cognitive Psychology. arXiv: 2210. 03737.

Vinanzi, Samuele, Angelo Cangelosi, and Christian Goerick. 2021. "The Collaborative Mind: Intention Reading and Trust in Human–Robot Interaction." *Iscience* 24 (2): 102130. https://doi.org/10.1016/j.isci.2021.102130.

Vinanzi, Samuele, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. 2019. "Would a Robot Trust You? Developmental Robotics Model of Trust and Theory of Mind." *Philosophical Transactions of the Royal Society B* 374 (1771): 20180032. https://doi.org/10.1098/rstb.2018. 0032.

Wagner, Alan R., Paul Robinette, and Ayanna Howard. 2018. "Modeling the Human–robot Trust Phenomenon: A Conceptual Framework Based on Risk." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8 (4): 1–24. https://doi.org/10.1145/3152890.

Appendix

A.1. Table version 1

Figure A1. First version of the table, which was evaluated in the first phase of evaluation. It includes an extensive distinction among all the five possible roles each teammate can play when performing a task in a team composed of one human (H) and one machine (M). The table assesses the feasibility of each teammate individually, through their dimensions of competence (C), possibility (P, which included the external factors), and intention (I). Each fillable cell (the dimensions) could be filled with green (g) or red (r) colours. This automatically fills in the overall feasibility of the teammate for that role (M column for machine's feasibility and H for human's). The feasibility columns can be (1) green if all three dimensions are green, (2) yellow if intention is red and all others are green, (3) orange if competence is red, or (4) red if another combination of red occurs. The final feasibility column can be (1) green if both feasibilities are green, (2) yellow if one is yellow and the other is green or yellow, (3) orange if one is orange and the other is green or orange, and (4) red if another combination occurs. Grey is ignored when calculating feasibility.