The Director Task: a Psychology-Inspired Task to Assess Cognitive and Interactive Robot Architectures

Guillaume Sarthou, Amandine Mayima, Guilhem Buisan, Kathleen Belhassein, Aurélie Clodic

The Director Task: a Psychology-Inspired Task to Assess Cognitive and Interactive Robot Architectures

Guillaume Sarthou[1], Amandine Mayima[1], Guilhem Buisan[1], Kathleen Belhassein[1,2], and Aurélie Clodic[1,3]

***Abstract*— Assessing robotic architecture for Human-Robot Interaction can be challenging due to the number of features a robot has to endow to perform an acceptable interaction. While everyday-inspired tasks are interesting as reflecting a realistic use of such robots, they often contain a lot of unknown and uncontrolled conditions and specific robot behavior can be hard to test. In this paper, we propose a new psychology-inspired task, gathering perspective-taking, planning, knowledge representation with theory of mind, manipulation, and communication. Along with a precise description of the task allowing its replication, we present a cognitive robot architecture able to perform it in its nominal cases. We finally suggest some challenges and evaluations for the Human-Robot Interaction research community, all derived from this easy-to-replicate task.**

I. Introduction

Developing a robotic architecture adapted to Human-Robot Interaction (HRI) and thus able to carry out interaction in an acceptable way is still today a real challenge. Such architecture should provide the robot with the capability of perceiving its environment and its partners, of interpreting this information, of communicating about it, of planning tasks with its partner, of estimating the others' perspective and mental state, etc. Once developed, evaluating these architectures can be difficult. The tasks we want the robot to handle must highlight a maximum of these abilities, while still being simple enough to be reproduced and to allow to conduct user studies.

For years, many tasks and scenarios have been inspired by everyday activities with the disadvantage of not highlighting some subtle abilities, but necessary for good interaction. The robot guide task [1] in mall, museum, or airport, requires high communication skills to understand queries and respond to them, whether to indicate a direction or to give advice. However, the perception needs are limited due to the vast environments, as well as the perspective-taking needs due to the same perception of the environment by the robot and the human. Finally, the human partner is not an actor of the task and just has to listen to the robot once their question is asked. Even if being in smaller environments, bartender-like tasks have the same disadvantages [2]. Indeed, the human is considered as a customer, and as such, the interaction with the robot is limited. The robot will never ask the human to help it for performing a task and their actions do not need to be coordinated.

Assembly tasks [3] aim at involving the human partner in the action of the task, requiring him to act with the robot. Nevertheless, the human acts as an assistant rather than as a partner. The robot thus asks for help when detecting errors (e.g., when it cannot reach some pieces), leading to unidirectional communication. Moreover, there is no belief divergence and no need for perspective-taking as both the robot and the human have the same knowledge about the environment.

Scaling down the tasks and adapting them to table-top scenarios allow to make the robot and the human to work in the vicinity of each other. In the assembly task presented in [4], the human is more involved in the task as they ask the robot to take pieces and to hold them to help them assembling a chair. Communications are unidirectional from the human to the robot but imply objects referring with the use of various visual features. Even if both agents have the same knowledge about the environment, the communication is grounded according to the current state of the world. However, no decision has to be made by the robot.

To study especially the perspective-taking ability and the belief management, the Sally and Anne scenario, coming from a psychology test, has been studied in robotic [5]. While the task is interesting to highlight these abilities, the humans do not have to act with the robot, and no communication is needed. The robot is only a spectator of the scene and no goal is formulated.

In section II, we first propose a new psychology-inspired task that is challenging for the Human-Robot Interaction community and rich enough to be extended. Inter alia, it requires perspective-taking, planning, knowledge representation with theory of mind, manipulation, communication, and decision-making. In section III, we then present a cognitive architecture able to perform the task in its nominal cases. Finally, in section IV, we present a discussion about the possible challenges and evaluations for the research community, resulting from possible extensions of this task.

As part of the Open Science movement, supported by the European Union[1], the implementation of the cognitive architecture we present is available in open-source[2]. Therefore, we encourage other researchers to reproduce our experiments. Furthermore, these software can serve as an inspiration or a base for the ones who wish to take up the challenges we propose.

[1]LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France firstname.surname@laas.fr
[2] CLLE, Université de Toulouse, CNRS, UT2J, Toulouse, France
[3]Artificial and Natural Intelligence Toulouse Institute (ANITI)

[1]https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-access_en

[2]https://github.com/LAAS-HRI/dt_laas_architecture

II. The Director Task: From psychology to HRI

A. The Director Task as used in psychology

The Director Task has been mainly used in psychology researches as a test of the Theory-of-Mind usage in referential communication. This task originates from a referential communication game from [6]. It was then adapted by Keysar and collaborators [7] to study the influence of mutual knowledge in language comprehension by using eye-tracking, becoming the Director Task. In this task, two people are placed one in front of the other with a vertical grid composed of different cells between them. The **director**, a participant or in most cases an accomplice, instructs the **receiver**, a participant, about objects to move in the grid. The particularity of this task is that the director, being on the other side of this grid, does not have the same perspective as the participant, thus does not see some cells that are hidden from their perspective. Then for a successful performance, participants must take the perspective of the director and update it all along the interaction.

For example in Fig. 1, if the director asks for the smallest apple (*), the proper smallest (called competitor) is only visible by the participant and not by the director. The participant then must understand the director's perspective to take the target apple and not the competitor one. Some studies showed that for their first attempt, participants considered or took the smallest apple from their own point of view and only after, the target one. These results were interpreted in [8], [9], [10], [11] as the participants understanding language in an egocentric way. Some social cognition studies used a computer-version of the Director Task [12] whose results are consistent with the ones mentioned previously, namely that participants do not use Theory-of-Mind inferences in language interpretation.

Although they require the attribution of mental states to others, some authors have distinguished Theory-of-Mind tasks and perspective-taking tasks reporting distinct although related mechanisms. In [13], they considered in their study that perspective-taking abilities were measured by the Director Task whereas Theory-of-Mind usage was investigated through another task called "strange stories" [14]. However, this Theory-of-Mind task requires the attribution of mental states to a story protagonist (to have knowledge of others' mental states), whereas the Director Task asks for adopting the perspective of the director in order to follow their instructions (to use this knowledge in order to execute the task properly). Thus, the authors estimated that the Director Task requires a higher degree of self-other distinction by continuously isolating our own perspective from the director one. In addition to perspective-taking abilities, the Director Task makes use of executive functions [15] and attentional resources [16].

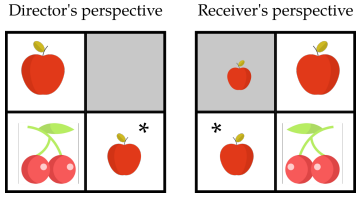

Fig. 1. Sample display from the director's and the receiver's perspectives. The asterisk indicates the target object / competitor.

The Director Task has thus been particularly used in psychology studies of referential communication, language comprehension, and perspective-taking abilities. However, to date it has never been exploited in the context of a HRI although this task presents interesting challenges for this field. It would not only bring technical challenges but also provide a way to investigate the different cognitive and behavioral processes involved in such a cooperative Human-Robot task.

B. The Director Task adaptation for HRI

In this section, we present the DT-HRI, the Director Task as we designed it for HRI, keeping the principle of two participants with a vertical grid between them. The high-level goal of the task is known by both agents: to put a set of blocks away. The precise goal is given by the experimenter to the director, either the robot or the human, i.e., the set of blocks that the receiver should remove from the compartments (see Fig. 2).

As mentioned in the previous section, the Director Task characteristics bring a number of interesting challenges for a collaborative robot to solve. Because this is a task with two roles, one of the first challenges is to build a robotic architecture that gives the robot the ability to play both roles. Then, each role brings some problems to solve from a robotic point of view. In the original task, the director knows they have a subset of the receiver's perspective, they can consider all the objects when communicating. Thus, only the receiver has to reason about the other's perspective, taking into account that some objects are not visible by the director. In order to enrich the task for HRI application, we propose to also have compartments hidden from the receiver and visible by the director (see Fig. 2). Therefore, both roles have to perform perspective-taking, whether to give instructions or to understand them. On one hand, this challenging task allows to demonstrate the abilities of a robotic system. On the other hand, it is an easily reproducible scenario to perform user studies on human-robot interactions in a controlled environment.

To be able to study more specifically some skills, such as verbal communication, perspective-taking and adaptation, we defined a set of rules for both the robot and the participant. First, to focus the task on verbal communication, the agents are **not allowed to point** to objects, either with their hand or gaze. Then, to strengthen the perspective-taking aspect and not fall into a simple referential communication task, participants are **not allowed to use geometrical relations** in the verbal communications. They cannot, for example, say "the leftmost block" or "the block to the right of the green one". In this way they are limited to few visual features, with high ambiguity, therefore requiring to take into account the other perspective. Finally, to enable an evolution of the

situation over time and thus requiring a constant adaptation during the interaction, the objects are not moved from one compartment to another but removed from the compartments. The **order of the instructions is free**, enabling the director to elaborate a strategy if needed.

1) A task to demonstrate the abilities of a robotic system:

a) Perspective-taking abilities: When working on the ToM in the HRI context, the Sally-Anne test has been used multiple times and allowed to demonstrate some systems [5]. But, one of the benefits of the Director Task compared to the Sally-Anne test is that the agents (human or robot/director and receiver) have not only to infer knowledge using the other's point of view but also to act so it is possible to acknowledge that they use it in decision-making.

b) Communication abilities: Moreover, the task requires to put a focus on communications which is widely studied in HRI. Indeed, the communication about an object can be more or less efficient, depending on the number of characteristics given about the object or the pertinence of these characteristics (e.g., in Fig. 1, the director does not need to add "red" to "take the small apple" as there is no apple of a different color). The robot needs to be able to give proper instructions but also to understand the human ones.

c) Planning abilities: When a large number of blocks has to be taken in the task goal, it quickly becomes complicated to communicate about some of them as the director would have to add a lot of adjectives to be able to refer to one block. Therefore when the robot is the director, it becomes interesting to integrate the communication and the task planning together. Indeed, depending on the order in which the blocks are designated, the complexity of instructions can decrease or increase. Then, the planner can return an optimal order in which the robot has to give the instructions to the human.

d) Contingencies handling abilities: While performing the Director Task, errors can happen. Either because the director gives a wrong instruction or the receiver misunderstands the instruction and takes the wrong block. In both cases, it can be because of a wrong consideration of the other agent's perspective. In the latter case, the instruction might be right but hard to interpret by the receiver leading to an error from them. Finally, errors can happen because of a failed action execution (e.g., a block falls on the floor), a system failure for the robot, inattention from the human, etc. A robot with a robust decision-making system will be able to analyze, try to determine their origin, and handle a number of these contingencies. For example, if the human takes the wrong block, the robot can react in different ways, e.g., asking the human to put it back or saying nothing and re-planning if this block was among the ones to take. If errors happen repeatedly, the robot can react differently than for a punctual error.

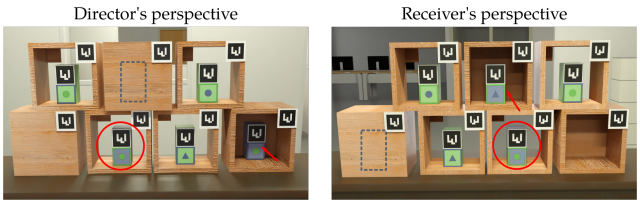

Fig. 2. A director task setup adapted to the HRI with the two perspectives. For the material, each element is equipped with AR-tags allowing their detection by the robot. Each block has four visual characteristics: a main color, a border color, a geometric figure, and a figure color. Compartments can be hidden for the director or the receiver. For the director to designate the block marked with a red circle, using the receiver's perspective, he can refer to it by its main color (blue) because the other blue block is not visible by the receiver. For the receiver, by taking into account the director's perspective, he can understand the referred block as the other blue block is not visible by the director.

2) A task to perform user studies: The easy replication and control of the setup make it a good task for human-robot user studies. Moreover, as it involves perspective-taking, communication, planning, and errors, there are a lot of elements that can be analyzed and explored. Also, with the same setup, it is possible to perform human-human studies or human-robot studies which can be interesting to compare.

C. A simple material

The material used in this task has been chosen to be easily acquired and can be hand-built. It is composed of twelve blocks, twelve compartments, and one storage area, each equipped with AR-tags allowing the robot to perceive them without advanced perception algorithms.

As shown in Fig. 2, the blocks have a primary color covering them all. On two opposite faces, additional visual features are drawn. The top part of these faces is dedicated to the robot's perception with unique AR-tag on each face[3]. The bottom part is the same on both faces and is dedicated to the human perception with a main color, a border, and a geometric figure. Every visual feature (the colors and the forms) has exactly two variants. The colors are either blue or green and the figures are either a triangle or a circle. The figures and colors have been chosen in such a way to allow the emergence of "coded words" between the participant to identify a block. With a bit of imagination, some could refer to the left-most block through the sentence "the mountain in the sea" or the second leftmost by "the puddle". The number of features has been chosen to have sixteen block variants from which we remove the four uni-color variants (all the elements having the same color) to avoid too easy description of the kind "the fully green block". Regarding their description complexity, while the main color is directly related to a block, the other colors are respectively related to the border and the figure. This means that for two blocks whose only difference is the color of one of these elements, the said element has to be referred to by its color. A description of a block involving all its features would be "the [color] block with the [color] border and the [color] [figure]". Such complete descriptions are hard for the human to process. In this way we expect the participants to minimize the complexity of their communication by referring to the

[3]because the tags are different on each side, the director can not refer to them as the receiver does not see the same ones

blocks only using the features distinguishing them from other blocks.

Three types of compartment exist. Some are open on two of their opposite sides allowing both the receiver and director to see the content and to manipulate it. Some are open only on one of their sides meaning that only one of the participants can see and take what is inside. The other participant can thus neither know if a block is inside or not. The last compartment type has an open side and the opposite one equipped with a wire mesh. Because of the side with the wire mesh, both participants can see what is inside but only one of them can take it. With these three types, we will be able to test the impact of the awareness of the blocks (e.g., a block is known to be present but not necessarily visible), the visibility of the blocks, and their reachability (e.g., a block can be visible but not reachable).

Finally, one storage area, corresponding to the place where the receiver has to store the blocks, is delimited by a rectangle on a shelf.

III. The Cognitive Robot Architecture

In this section, we present DACOBOT (Deliberative Architecture for COllaborative roBOT), the architecture developed to handle the Director Task in its nominal case[4] but also to allow for future extensions, endowing the robot with the abilities described in section II-B.1. This architecture is a whole new instantiation of the deliberative architecture for Human-Robot Interaction presented in [17]. The seven identified modules are represented in Fig. 3 with their respective communication links. In the rest of the section, we detail each module and how we have refined them in terms of functionality and linking.

A. Storing and reasoning on symbolic statements

The knowledge representation is always a core component of cognitive architectures as this knowledge allows the robot to understand the environment it evolves in. Moreover, this same knowledge makes the robot able to communicate with its human partner about the current state of the world and ground the partner's utterance regarding this same world state.

Some have chosen to propagate their knowledge all along their architecture [18], each component enriching this knowledge at each stage. Others have preferred to see their knowledge base as an active server activating perception process regarding the searched information when needed [19].

As the architecture on which we based ours, we chose a central, server-based, knowledge base. We however refined it into two distinct sub-modules, the semantic knowledge base and the episodic one. The semantic part is in charge of representing the environment elements meaning, the objects' and agents' types, their applicable properties, the descriptions and parameters of the actions, a part of the language model with verbs or pronouns, and their names in natural language. Besides, it is also in charge of representing

[4] https://www.youtube.com/watch?v=jtSyZeqBkp0

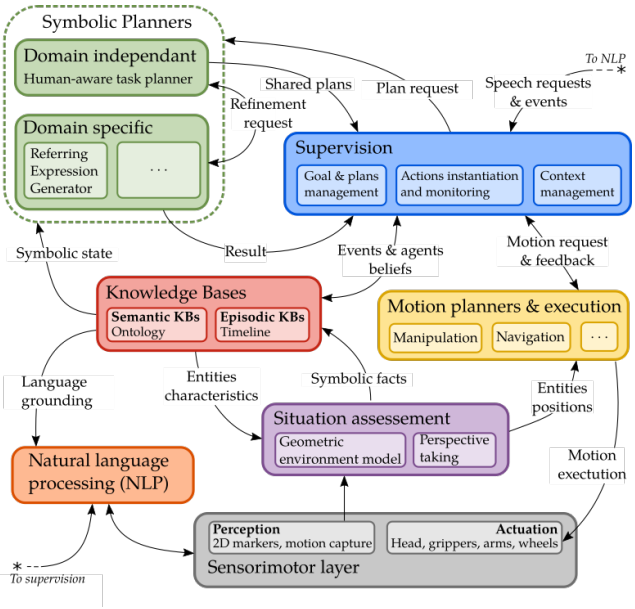

Fig. 3. An overview of the DACOBOT architecture developed to handle the Director Task. Each block does not necessarily represent one software component but rather an architectural module (in terms of the features it implements). The arrows represent the type of information exchanged between the modules.

the current symbolic world-state (the computed facts) and thus the instantiation of the concepts in terms of physical (e.g., this particular block) or abstract (e.g., this particular action instance) entities. Among these instantiations, we will have the blocks' visual features, their computer-aided design (CAD) models, or their tags ids. The episodic part aims at keeping a trace of the symbolic transitions of the world in the time. It is strongly linked to the semantic knowledge base as it allows to semantically interpret these transitions.

The semantic knowledge base is in the form of an ontology as it allows a rich expressiveness, a standardisation of the representation among several architectures, and reasoning capabilities. We chose the software Ontologenius [20] to manage it. A reason for this choice is that it is fully adapted to HRI applications by representing the robot's knowledge and the estimation of the partners' knowledge separately, which refers to the psychological concept of the "self-other distinction" as coined in joint action study [21]. The episodic knowledge base has been developed on the same principle by representing one timeline per agent. Each timeline is thus connected to an ontology instance and reflects a specific agent perception of the evolution of the environment. Moreover, the episodic sub-module provides an event subscription mechanism taking advantage of the link with the semantic sub-module. A component can thus ask to be informed when a block is put on a table regardless of the specific block and table but above all, regardless if this event has been perceived or deduced through inferences.

B. Assessing the world: from geometry to symbolism

The role of the geometrical Situation Assessment module is first to gather different perceptual information and build

an internal geometric representation of the world. From this world representation, the module then runs reasoners to interpret it in terms of symbolic statements between the objects themselves and between the involved agents and the objects. Doing so, the module only builds the robot's representation that does not necessarily reflect what the human partner believes about the world. This is the case with the occluded compartments. If a block is present in a compartment occluded from the human perspective, this block is not visible and thus unknown to the human and should not exist in their representation of the world. Here is the second role of our Situation Assessment module, estimating the human's perspective and building an estimation of their world representation. It is the first step allowing to implement the theory of mind principles [22].

To implement this module, we have chosen the Underworld framework [23]. Its advantage is to not be monolithic. Its principle is to create a set of worlds, each working at a different granularity and integrating specific features. It allows easy reuse of existing modules and makes the core reasoning capabilities independent of the used perception modalities. The worlds' structure we use is represented in Fig. 4. At the top, there are the perception modalities, here AR-tags [24] for the objects and motion capture (mocap) system for the human detection. For each perception system, we define a world. In these worlds, we can filter the perception data depending on the system used. For the mocap, the data is clean enough. For the AR-tags we apply first a motion filter to discard data acquired when the robot moves and a field of view (FOV) filter to discard data from the border of the camera because of distortions. Moreover, both perception worlds can use the knowledge base presented previously to get the entities' CAD models and unique identifiers (UIDs) shared across all the components of the architecture. When the AR-tags world receives an AR id, it can query the semantic knowledge base to get the UID related to this tag and get its CAD model. As the output of these worlds, we ensure to have clean data with UID related to the knowledge base.

The world of the middle in Fig. 4 is the robot's world representation. Information from the perception worlds is merged along with the static elements (the building walls) and the robot model. From there, geometric reasoners are applied to extract symbolic facts. In the current version of the system, the computed facts are *isOnTopOf*, for an object put on top of another, *isInside*, for a block in a compartment, *isVisibleBy*, assessing if an agent could see the object or not, and *isReachableBy*, assessing if an object can be taken by an agent. All these facts are sent to the robot's semantic knowledge base, where reasoners will deduce further facts. For example, if a block is in a compartment, the compartment has the block inside (inverse property), and if this compartment is on top of the table, the block inside is also above the table (chain axiom).

While the previous world corresponds to the robot's representation, the one below aims at representing the partner's one. From the previous world, we compute a segmentation image from the human point of view and use it as a filtered perception world. This allows us to instantiate the same kind of world management process we used for the robot but this time for the human. In this way, we emulate their perception capability and the geometric reasoners can be run in the same way as previously. Symbolic facts are thus computed and sent to the human's semantic knowledge base. In the world of the bottom on Fig. 4, we can see that the two blocks in the occluded compartments are not present in the human world. Here we make explicit the difference between an object that is unknown and an object that is known but not visible.

C. Planning with symbolic facts

The symbolic planners are divided into two categories: the domain-independent, planning high-level tasks, and the domain-dependant, specialized in solving precise problems. We first introduce the domain-specific ones and the domain-independent in a second time.

1) Solving precise problems: Building a single monolithic planner could be an intractable challenge. Thus, we chose to consider a set of dedicated planners which could be reused from one system to another. In the current version of the system, only one specific planner has been identified. This planner is a Referring Expression Generation (REG) solver. Regarding the current symbolic state of the world, it aims at finding the minimal set of relations to communicate and allow the listener to identify a given entity. For example, wanting to refer to a block being the only one with a green triangle on it among the other, this planner can find that the

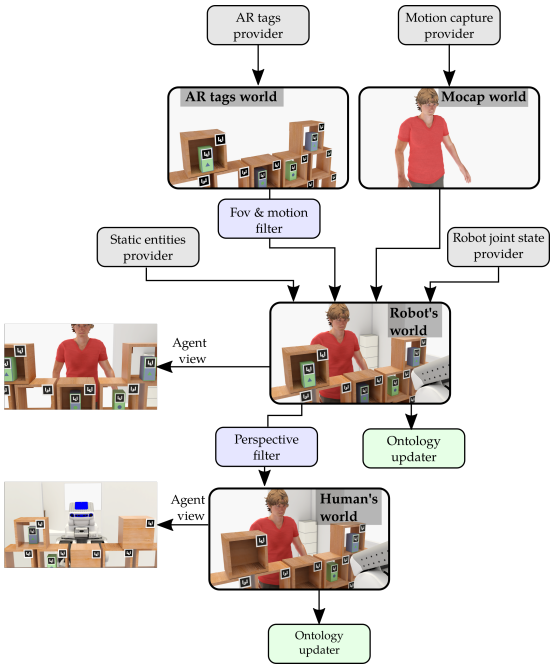

Fig. 4. The world cascading structure of the geometrical situation assessment system. The two worlds at the top are build from the perception systems and filtered. The world of the middle merges the different perception information and computes symbolic facts on it. The world at the bottom is the estimation of the human world representation and is computed based on perspective-taking in the robot's world. Like for the world of the middle, symbolic facts are computed and sent to the semantic knowledge base.

only relations to communicate are the block's figure and the figure's color. With this information, the listener should be able to identify the referred block without ambiguity.

This planner is presented in [25] which is based on a Uniform Cost Search algorithm and which is to the date the most efficient one in term of computation time. It works with an ontology, being the semantic knowledge base presented previously. Because the communication information it generates will be interpreted by the robot's partner, we chose to give the estimated human knowledge base as input to the planner. Thanks to this, the blocks unknown from the human — i.e., hidden from them — are not taken into account as they cannot lead to any ambiguity for the listener. Moreover, this planner can take some constraints as input, related to the property usability and the context of the communication. The usable properties constraint prevents some properties to be used in a referring expression. Indeed, the input ontology is not dedicated to the specific referring expression generation problem and contains additional knowledge used by other modules as the objects' CAD models or tag UIDs, that does not aim to be communicated. The communication context aims at representing relations assumed to be already known by the listener. For the Director Task, when the robot asks the human to take a block, it assumes they know it is only talking about objects above the table around which the robot and the human are interacting. The already stored blocks — not on the table anymore — are thus not taken into account in the communication. If needed the communication context can be refined, for example by defining that the robot — and thus should the human as well — will only consider visible blocks and reachable blocks.

2) Planning for self and others: In the context of a Human-Robot interaction, when planning how to perform a high-level task, one has to take into account the human's contribution. Our current task planner is thus based on the principle of the Human-Aware Task Planner (HATP) [26]. The latter can generate a shared plan in which parts of the task are assigned to the human partner, depending on some criteria. However, the robot's partner is not an agent that the planner can directly control. Indeed, it must sometimes communicate about the plan to inform the human about their next actions. This is why our current task planner [27] allows the robot to plan by emulating the human decision, action, and reaction processes. For the Director Task, emulating the human reaction to a given instruction enables the comparison between multiple blocks order, the communication of higher-level instructions to the human (e.g., ask to withdraw rather than take then put down) and the balance between multiple communication modalities.

As at execution domain-specific planners are used, and because the task planner uses the same type of knowledge representation, it can use them during the planning process. In the current architecture, it can thus estimate the cost and the feasibility of referring communication by calling the REG. This method previously implemented with HATP [28] has been successfully integrated with the new planner.

D. Managing the interaction

JAHRVIS (Joint Action-based Human-aware supeRVISor) constitutes the decisional kernel of this cognitive architecture. Like its predecessors, SHARY [29] and its extensions [30], [31], it is a supervision system for a human-aware robot, handling not only the robot's action execution but also the estimation of the human mental state, the monitoring of the human actions, and the communication with them. These features are implemented through several processes:

a) Interaction sessions management: JAHRVIS is designed to handle an interaction at three levels: *interaction sessions*, *tasks (here the Director Task)* and *actions*. We defined an *interaction session* as the period during which the robot and a human interact together and are engaged. It is divided in three parts : the *greetings*, the *body of the interaction* and the *goodbyes*. Therefore, the interaction session management process is in charge of the exchanges happening at the beginning of interaction (the greetings), at the end (the goodbyes) and all events/exchanges happening outside tasks (e.g., conversation, goal negotiation) or during a task but not related to it (e.g., human leaving whereas it is not planned in the task, human doing a parallel task on its own).

b) Communication management: Human-Robot Interaction involves verbal and non-verbal communication. Communications from an agent to another are categorized as follows in JAHRVIS: to give information which leads to a belief update for the agent receiving the communication; to ask a question which is normally followed by an answer leading to a belief update for the agent who first asked the question; to ask the other agent to perform an action leading to a belief update for this agent and to the execution of the given action; conversation, i.e., all dialog not related to a task or a goal/plan negotiation.

c) Human management: As JAHRVIS is dedicated to HRI, it is essential that it has a process managing the human beliefs about the interaction session, the ongoing task, and plan. It gathers the data provided by the other JAHRVIS processes and the other modules of the architecture. It makes sure that the human has all the knowledge they need for what they have to perform and if not, it hence acts or communicates through the other processes.

d) Task management: This process ensures the monitoring and the execution of the plan of the task on which the human and the robot agreed to perform. When a plan contingency happens, it can react based on its knowledge about the human and the environment and perform a repair thanks to action or communication.

e) Quality of Interaction management: All along an interaction session and a task, JAHRVIS estimates in real-time the Quality of Interaction (QoI) [32] based on data from the other processes. This evaluation process is focused on two elements: the measure of human engagement and the measure of the effectiveness of collaborative task performance. It thus provides additional information to the decision-making process (at the task or the interaction session management level) and opens the possibility for

reconsidering the robot's behaviour in case it estimates that the Quality of the Interaction is degrading (e.g., changing its plan or the way it is achieving it, changing the cost or a constraint in one of the planner, informing the human or requesting a change in their behaviour, or even deciding to disengage).

IV. Open challenges for the community

So far, we proposed a cognitive robot architecture handling the Director Task in its simplest form, both roles. In this section, we now present some open challenges for the community around the task. Moreover, because the task can be performed in a controlled environment, we also present in a second part some user studies to investigate ways of sharing information.

A. Some challenges to take up

Challenged abilities / components	Challenges
Perspective-taking	1
Communication	4, 6, 7
Task planning	2, 3, 4
Reference generation	4, 5, 8, 9
Contingencies handling	1, 2, 3, 4

1) Fine contingency analysis: Due to the high ambiguity between the blocks and the presence of occluded compartments, failures can easily arise and have to be handled. In the case the human is the receiver and does not take the instructed block, the robot has to determine the origin of the failure. It could come from a perspective not taken into account either by the director or the receiver, a block description not clear enough, or just an error of the receiver regarding a correct (non-ambiguous) description.
2) Not handling contingencies as errors: Based on the example of the previous challenge, the human takes another block than the one instructed but this block could be part of the next ones to take in the plan. In this case, why the robot should try to "repair" i.e., make the human takes the instructed block? Maybe it could mention to them that they took the wrong block but it does not matter because this one is also part of the plan. Then, either the robot could re-plan or even better, use a conditional plan and adapt according to the human's actions.
3) Handling errors as errors: Still based on the case where the human takes another block than the one instructed, the robot has to communicate and negotiate with them in order, first to fix the error i.e., put back the block to its original compartment, then adapt its original instruction to make it clearer and improve the chances to have them take the right one.
4) Changing something when recurrent failures: In case of recurrent failures by the partner, during one interaction session (multiple tasks can be performed in one session) or along with several ones, the robot could try to analyse the origin of the failures and adapt itself to increase the QoI and reduce the failures. It could be through properties' cost adaptation if the partner has some difficulties with certain visual features or communication context adaptation if the partner took the stored blocks into account in its understanding.
5) Allowing spatial references: As explained in section II-A, the Director Task is originally a task to test referential communication. Even if the present version asks the participants to not use spatial reference, this rule could be relaxed to study perspective-corrected spatial Referring Expression Generation.
6) Understanding the human instructions: In the current implemented version, the robot can only understand a precise vocabulary, being the one describing the blocks in the way we have thought them. In a more natural interaction, humans could use a richer vocabulary, give instruction in multiple steps or have communications not directly linked to the task. During tests for designing the task, it was common to have instructions like "take the block with a ... triangle. No, rather the one with a green border". Such complex communications should have to be managed by the robot.
7) Introducing code words: As presented in section II-C, the visual features on the blocks have been designed to be able to see landscapes on them, with a little imagination. During the interaction, the robot could thus try to negotiate some coded words in order to be more efficient in the task considering multiple sessions. Being the receiver, it would have to understand the coded words as to be part of a description and remember them.
8) Referring to a past event: When a human performs multiple times the Director Task with the robot, noteworthy events can happen. These events could be recognized and recorded by the robot so it can refer to them when speaking about an object (e.g., "can you take the one you dropped in the previous task ?"). Likewise, the human may also use these past events and the robot would have to understand them.
9) Communicating about multiple blocks in a raw: Instead of giving instructions one at a time, the director could give instructions for multiple blocks in a raw. This may bring different kinds of communications from the base task as "I do not remember the instruction for the last block" when the human is the receiver. Also, this would be interesting for planning when the robot is the director as it could give instructions such as "Take all the blocks with a triangle on them" and it would be a different kind of instructions to interpret when the robot is the receiver.

B. Some Director Task-based user studies to perform

Some robot behaviours, mainly about the referring expression generation, have been designed with regard to the current literature but could be refined thanks to user studies based on the Director Task. The references to the blocks involve the minimum of visual features allowing to discriminate them without ambiguity to fit the Grice's Maxim of Quantity [33]. However, due to all the cognitive mech-

anisms to use in this task (e.g., perspective-taking) and the high ambiguity among the blocks, evaluating such behaviour compared to a full explanation could be interesting. Indeed, giving a reference with more information than needed would ensure to not match blocks being only visible by the receiver, which could help them to select the right block.

As presented in section II-C, a special compartment equipped with a wire mesh can be used. Referring to a block matching also the one in this particular compartment could disturb the receiver or at least require a higher cognitive load to determine the right block to take. Such behavior could also be interesting to evaluate. In the same way, a block that was visible by the receiver and that the director move in a hidden compartment could disturb the receiver.

Evaluating such behaviours in a controlled task where the participants cannot know the real goal of the study could help the community in the design of architectures applied to more realistic scenarios.

V. Conclusion

This article presents a new task inspired by psychology to assess cognitive robot architecture capacities, highlight them and challenge them: the Director Task. This task involves cognitive abilities such as perspective-taking, communication, planning, and contingency handling which are studied a lot in HRI. Along with this task, we describe the cognitive architecture we built. It is currently able to perform the task in the nominal case with the robot being the receiver or the director.

The Director Task we propose aims to be extended and be a sandbox for the HRI community. We have presented nine possible challenges to be taken up and two possible user studies to be carried out. As a base to be reused, the components of the architecture are released in open-source to anyone who would like to pick few components or to use it in its integrity. From there, new features can be implemented, improving the architecture abilities.

Acknowledgement: Authors want to thank Yannick Riou and Alexandre Bonneau for their work on the robotic integration. This work has been funded by the Agence Nationale de la Recherche JointAction4HRI project ANR-16-CE33-0017 and the Artificial and Natural Intelligence Toulouse Institute (ANITI).

References

[1] S. Satake, K. Nakatani, K. Hayashi, T. Kanda, and M. Imai, "What should we know to develop an information robot?" *PeerJ Computer Science*, 2015.

[2] R. P. Petrick, M. E. Foster, and A. Isard, "Social state recognition and knowledge-level planning for human-robot interaction in a bartender domain," in *AAAI Workshop on Grounding Language for Physical Systems*, 2012.

[3] S. Tellex, R. Knepper, A. Li, D. Rus, and N. Roy, "Asking for help using inverse semantics," 2014.

[4] J. Brawer, O. Mangin, A. Roncone, S. Widder, and B. Scassellati, "Situated human–robot collaboration: predicting intent from grounded natural language," in *IEEE/RSJ IROS*, 2018.

[5] G. Milliez, M. Warnier, A. Clodic, and R. Alami, "A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management," in *IEEE RO-MAN*, 2014.

[6] R. M. Krauss and S. Glucksberg, "Social and nonsocial speech," *Scientific American*, 1977.

[7] B. Keysar, D. J. Barr, J. A. Balin, and J. S. Brauner, "Taking perspective in conversation: The role of mutual knoowledge in comprehension," *Psychological Science*, 2000.

[8] B. Keysar, "The illusory transparency of intention: linguistic perspective taking in text," *Cognitive Psychology*, 1994.

[9] B. Keysar, D. J. Barr, and W. Horton, "The egocentric basis of language use: insights from a processing approach," *Current Directions in Psychological Sciences*, 1998.

[10] B. Keysar and D. J. Barr, "Self-Anchoring in Conversation: Why Language Users Do Not Do What They "Should"," in *Heuristics and Biases*. Cambridge University Press, 2002.

[11] B. Keysar, S. Lin, and D. J. Barr, "Limits on theory of mind use in adults," *Cognition*, 2003.

[12] I. Dumontheil, I. A. Apperly, and S.-J. Blakemore, "Online usage of theory of mind continues to develop in late adolescence," *Developmental science*, 2010.

[13] I. Santiesteban, S. White, J. Cook, S. J. Gilbert, C. Heyes, and G. Bird, "Training social cognition: from imitation to theory of mind," *Cognition*, 2012.

[14] F. G. E. Happé, "An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults," *Journal of Autism and Developmental Disorders*, 1994.

[15] P. Rubio-Fernández, "The director task: A test of Theory-of-Mind use or selective attention?" *Psychonomic Bulletin & Review*, 2017.

[16] S. Lin, B. Keysar, and N. Epley, "Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention," *Journal of Experimental Social Psychology*, no. 46, 2010.

[17] S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic, and R. Alami, "Artificial cognition for social human–robot interaction: An implementation," *Artificial Intelligence*, vol. 247, 2017.

[18] N. Hawes, M. Zillich, and J. Wyatt, "Balt & cast: Middleware for cognitive robotics," in *IEEE RO-MAN*, 2007.

[19] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. K. Bozcuoğlu, and G. Bartels, "Know rob 2.0—a 2nd generation knowledge processing framework for cognition-enabled robotic agents," in *IEEE ICRA*, 2018.

[20] G. Sarthou, A. Clodic, and R. Alami, "Ontologenius : A long-term semantic memory for robotic agents," in *IEEE RO-MAN*, 2019.

[21] E. Pacherie, "The phenomenology of joint action: Self-agency vs. joint-agency," in *Joint Attention: New Developments*, S. Axel, Ed. MIT Press, 2012.

[22] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind" ?" *Cognition*, 1985.

[23] S. Lemaignan, Y. Sallami, C. Wallhridge, A. Clodic, T. Belpaeme, and R. Alami, "Underworlds: Cascading situation assessment for robots," in *IEEE/RSJ IROS*, 2018.

[24] M. Fiala, "Artag, a fiducial marker system using digital techniques," in *IEEE (CVPR'05*, vol. 2, 2005.

[25] G. Buisan, G. Sarthou, A. Bit-Monnot, A. Clodic, and R. Alami, "Efficient, situated and ontology based referring expression generation for human-robot collaboration," in *IEEE RO-MAN*, 2020.

[26] R. Lallement, L. De Silva, and R. Alami, "HATP: An HTN Planner for Robotics," in *ICAPS Workshop on Planning and Robotics*, 2014.

[27] G. Buisan and R. Alami, "A human-aware task planner explicitly reasoning about human and robot decision, action and reaction," in *Companion of the ACM/IEEE HRI*, 2021.

[28] G. Buisan, G. Sarthou, and R. Alami, "Human aware task planning using verbal communication feasibility and costs," in *ICSR*, 2020.

[29] A. Clodic, H. Cao, S. Alili, V. Montreuil, R. Alami, and R. Chatila, "Shary: A supervision system adapted to human-robot interaction," in *Experimental Robotics*, O. Khatib, V. Kumar, and G. J. Pappas, Eds. Springer Berlin Heidelberg, 2009.

[30] M. Fiore, A. Clodic, and R. Alami, "On Planning and Task achievement Modalities for Human-Robot Collaboration," in *ISER*, 2014.

[31] S. Devin and R. Alami, "An Implemented Theory of Mind to Improve Human-Robot Shared Plans Execution," in *ACM/IEEE HRI*, 2016.

[32] A. Mayima, A. Clodic, and R. Alami, "Toward a Robot Computing an Online Estimation of the Quality of its Interaction with its Human Partner," in *IEEE RO-MAN*, 2020.

[33] H. P. Grice, "Logic and conversation," in *Speech acts*. Brill, 1975.