TU Delft

Delft University of Technology

SONAR

An Adaptive Control Architecture for Social Norm Aware Robots

Dell'Anna, Davide; Jamshidnejad, Anahita

DOI
10.1007/s12369-024-01172-8

Publication date
2024

Document Version
Final published version

Published in
International Journal of Social Robotics

Citation (APA)
Dell'Anna, D., & Jamshidnejad, A. (2024). SONAR: An Adaptive Control Architecture for Social Norm Aware Robots. *International Journal of Social Robotics*, *16*(9), 1969-2000. Article 105064. https://doi.org/10.1007/s12369-024-01172-8

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

International Journal of Social Robotics
https://doi.org/10.1007/s12369-024-01172-8

SONAR: An Adaptive Control Architecture for Social Norm Aware Robots

Davide Dell'Anna[1] · **Anahita Jamshidnejad**[2]

Accepted: 16 July 2024
© The Author(s) 2024

Abstract
Recent advances in robotics and artificial intelligence have made it necessary or desired for humans to get involved in interactions with social robots. A key factor for the human acceptance of these robots is their awareness of environmental and social norms. In this paper, we introduce SONAR (for SOcial Norm Aware Robots), a novel robot-agnostic control architecture aimed at enabling social agents to autonomously recognize, act upon, and learn over time social norms during interactions with humans. SONAR integrates several state-of-the-art theories and technologies, including the belief-desire-intention (BDI) model of reasoning and decision making for rational agents, fuzzy logic theory, and large language models, to support adaptive and norm-aware autonomous decision making. We demonstrate the feasibility and applicability of SONAR via real-life experiments involving human-robot interactions (HRI) using a Nao robot for scenarios of casual conversations between the robot and each participant. The results of our experiments show that our SONAR implementation can effectively and efficiently be used in HRI to provide the robot with environmental and social and norm awareness. Compared to a robot with no explicit social and norm awareness, introducing social and norm awareness via SONAR results in interactions that are perceived as more positive and enjoyable by humans, as well as in higher perceived trust in the social robot. Moreover, we investigate, via computer-based simulations, the extent to which SONAR can be used to learn and adapt to the social norms of different societies. The results of these simulations illustrate that SONAR can successfully learn adequate behaviors in a society from a relatively small amount of data. We publicly release the source code of SONAR, along with data and experiments logs.

Keywords Social norms · Social robots · Social norm-aware robots · Norm adaptation · Fuzzy logic · Belief-desire-intention · Large language models

1 Introduction

Recent advances in robotics and Artificial Intelligence (AI) make daily interactions with intelligent robots a close reality [1, 2]. A key factor for the acceptance of robots by humans is the social interaction and norm awareness of these robots [3]. Human behaviors and interactions are heavily regulated by social and personal norms [4, 5], which determine how people (should) behave in different situations, and improve their interactions by facilitating cooperation and communication [6]. The ability of the robots to understand and reason about human social norms improves the naturalness and effectiveness of the human-robot interaction and collaboration [7, 8]. In healthcare applications, for example, this implies higher chances for a patient to establish trust in an assistive robot, improving both the acceptance of the robot by the patient and the effectiveness of the therapeutic interventions [9].

Incorporating the norms within the real-time automated reasoning and decision-making of social robots requires approaches that can deal with the uncertainty, dynamics, and impreciseness of social norms.[1] [12, 13].

✉ Davide Dell'Anna
d.dellanna@uu.nl

Anahita Jamshidnejad
a.jamshidnejad@tudelft.nl

[1] Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

[2] Department of Control and Operations, Delft University of Technology, Delft, The Netherlands

[1] Norms are often stated and perceived by humans through vague and abstract linguistic terms [10]. Consider, as an example, the social norm concerning the *adequate distance with the other person during a*

Published online: 01 October 2024

In recent years, some practical approaches have begun to appear in the context of social and socially assistive robotics [12, 13] to overcome the limitations of traditional logics (e.g, deontic logic [14]), historically studied for normative reasoning of intelligent systems [15–17], but generally computationally intractable for real-time applications such as that of human-robot interaction [18–20]. For example, within the project CARESSES EU-Japan [21, 22], Bruno et al. [11] propose a framework for culture-aware robots based on fuzzy logic control. Fuzzy logic [23] is a non-traditional logic that allows to reason according to IF-THEN rules, whose core components are expressed via ambiguous and imprecise non-quantified *linguistic terms*. Fuzzy logic Inference Systems (FISs) use membership functions to specify the *degree* to which an element belongs to a fuzzy set of elements. For example, a *distance* of 2 meters can be considered as *medium* with a degree of 0.7 (on a scale [0, 1]), and at the same time as *low* with a degree of 0.5.

This characteristic of fuzzy logic makes it particularly suitable for automated inference and decision-making of robots in social contexts [24, 25], since the available and relevant knowledge (e.g., the preference, needs and background of a patient and the knowledge of the therapists) is often expressed via (fuzzy and ambiguous) linguistic terms. Preliminary studies [11, 26] have shown that fuzzy logic and fuzzy inference can effectively be used by social robots to autonomously reason about and to properly react to social norms such as proxemics based on cultural or individual preferences.

Existing works, however, are still preliminary and are mainly focused on specific case studies [8]. The state-of-the-art currently lacks a general framework for social robotics that supports high-level reasoning and decision-making while leveraging the practical advantages of fuzzy logic for modeling and reasoning about the norms. Moreover, a great majority of the existing works on normative reasoning does not consider *norm revision and adaptation*, despite their essence for dealing with (social) norms, which are intrinsically dynamic [4]. Norm revision and adaptation are currently an open challenge for computational normative systems [3, 27, 28].

In this paper, we introduce a novel adaptive control architecture: SONAR (for SOcial Norm Aware Robots). SONAR is a *general-purpose* and *robot-agnostic* architecture that leverages, on the one hand, the practical BDI (Belief-Desire-Intention) reasoning model [29] for high-level explainable [30, 31] automated decision-making of social robots, and on the other hand, fuzzy logic to provide adaptive norm-aware capabilities for these robots. We also contribute with a novel norm adaptation mechanism, based on the fuzzy context adaptation technique [32], for learning and adapting (the meaning of) social and personal norms at run-time, and for autonomously determining adequate behaviors in a society.

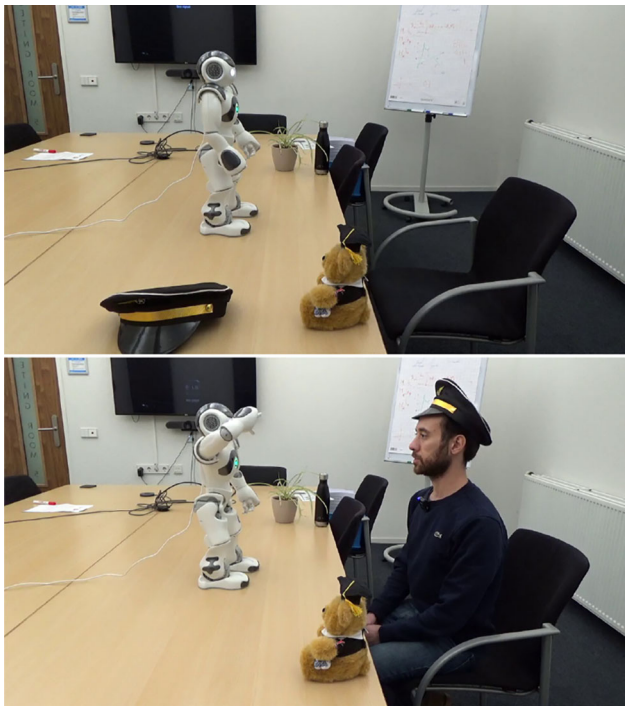


Fig. 1 Setup of the experiments with the Nao robot (top figure), and a snapshot of one experiment during a role-playing activity where the robot is expected to adapt to the norms within a hierarchical situation (bottom figure)

We run several exploratory experiments in the context of human-robot interaction using a Python 3.9 implementation of SONAR to steer the behavior of a NAO robot [33] in scenarios of casual human-robot conversations (see Fig. 1). Our experiments assess the feasibility and applicability of SONAR, and the perception of the human about various aspects (e.g., naturalness) of the social interaction of the robot, in comparison with an alternative robot that does not leverage social and normative reasoning, nor proactive behaviors. Additionally, we evaluate the proposed norm adaptation mechanisms by investigating, via computer-based simulation, the extent to which the robot can learn the social norms of different societies.

We publicly release the source code of SONAR and the results of our experiments, including an extensive data set of the corresponding human-robot interactions (see [34]). The data includes 50 conversations that occurred during our experiments between humans and Nao, where the answers of the robot were autonomously generated using a GPT-based large language model. The videos of the interactions are available upon request (via [35]).

footnote continued 1
conversation. The concept of *adequate distance* is not precisely defined and can vary depending on the particular social context, culture, and individual preferences [11].

The rest of this paper is structured as it follows. Section 2 provides a background discussion on related literature. Section 3 describes the proposed control architecture, SONAR, as well as the proposed mechanisms for norm adaptation in the course of human-robot interactions. Sections 4 and 5 represent, respectively, the setup and results of the experiments including real-life human-robot interactions. Section 6 reports on the evaluation of the proposed norm adaptation via computer-based simulations. Finally, Sect. 7 concludes the paper and proposes topics for future research.

2 Related Work

As a fundamental concept for coordinating human activities in societies [5, 36], (social) norms have been studied in a variety of different fields, including sociology [37, 38], philosophy [14, 39], economics [40, 41], AI [27, 28] and social robotics [7, 8, 10]. According to Castelfranchi et al. [42], in order to be considered *norm-aware*, autonomous agents, including social robots, should be able to recognize whether or not a norm exists for the given context, and to deliberately follow or violate these norms. Rato et al. [43], in line with Dignum et al. [44, 45], identify design principles for socio-cognitive systems to make them norm-aware. These include the capacity of the system to (i) construct a social context by ascribing social meaning to sensory information, (ii) adapt its behavior according to the social context, and (iii) attribute social categories to social actors. On the same lines, Castro et al. [46] discuss the following requirements for social agency of robots: (a) the behavior of a social agent must be rationally motivated by beliefs, desires, and intentions, (b) the agent must identify other agents and vary its behavior accordingly, (c) the agent must exhibit a tendency to engage in interactions, (d) the agent must be capable of understanding the behaviors of themselves and other agents, in terms of expectations generated by social norms, rules, and conventions, and should modify their behavior accordingly.

Among the decision-making models for intelligent systems in line with the requirements for social agency outlined above, the *belief-desire-intention* (BDI) reasoning model [29] has gained wide attention in AI and social simulation [47–50], leading to a variety of BDI-based architectures [51, 52] and (Agent Oriented) programming languages [53–55]. The BDI model implements the main aspects of Bratman's theory of human practical reasoning [29] by attributing to the agent mental states such as beliefs, desires, and intentions, and by characterizing the deliberation and reasoning of the agents in terms of these mental states [56]. Beliefs represent the informational state of the agent, i.e., beliefs about the world and rules of beliefs propagation (which beliefs can be derived from others). Desires (also often called goals) represent the motivational state of the agent, i.e., the objectives or situations that the agent would like to accomplish or bring about. Intentions represent the deliberative state of the agent, i.e., what the agent has chosen to do (has begun executing a plan). (Designing Buildings for Real Occupants: An Agent-Based Approach - Andrews) Castelfranchi [57] represent norms as mental objects that interact with beliefs, goals, and plans, and that impact the generation and selection of the goals and plans. Dignum et al. [58] discuss how to integrate deontic events as normative beliefs in BDI in the context of social agents.

BDI has been employed in social robotics in some preliminary studies, for example to add proactivity to robots (see [59–62]). The literature on social robotics that considers both BDI and social norms, however, is scarce. Among the few works, worth noting is that of Ribino et al. [63], where a framework similar to ours is presented, but specifically tailored for an indoor environmental quality monitoring case study.

Social norms in social robotics have been considered from many points of view. These include studies on social cues, such as robotic gaze responsiveness [64], the integration of affective computing techniques in robots [65], and studies on the effect of robot's visual appearance and robot's encouragement on people's perceptions and behavior [66, 67]. Recently, Kola et al. [68] suggested that the use of the DIAMONDS taxonomy of eight major dimensions of situation characteristics, proposed by Rauthmann et al. [69], can allow intelligent systems to perceive the social elements of a situation and to comprehend their meaning. Rauthmann et al. [69] analyzed the correlation between 30 different *situation cues*, i.e., physical and objective elements of a situation (e.g., who is present in a situation, what activity is taking place, etc.), and the 8 DIAMONDS *situation characteristics* (i.e., Duty, Intellect, Adversity, Mating, Positivity, Negativity, Deception, and Sociality) that represent *social and psychological meanings of situations*, for the two different societies of United States and Austria. They report, for example, that in the Austrian sample, duty had a positive correlation with the "working, studying" situation cue (with a correlation coefficient of 0.60) and a negative correlation with "TV, movies" (with a correlation coefficient of 0.31). Social behaviors of robots have also been studied in the context of social planning [70] and in healthcare contexts [71].

Despite the numerous works, many challenges still exist, especially concerning normative reasoning and representation [8]. In a recent survey, Avelino et al. [8] highlight that most existing works present a fixed pipeline of modules tailored for specific applications, and indicate that representation and learning of social norms is still an open challenge as many approaches do not support an explicit way to incorporate new norms.

Among the exceptions, Carlucci et al. [72] propose the use of Petri-nets to represent social norms explicitly. Wasik et al.

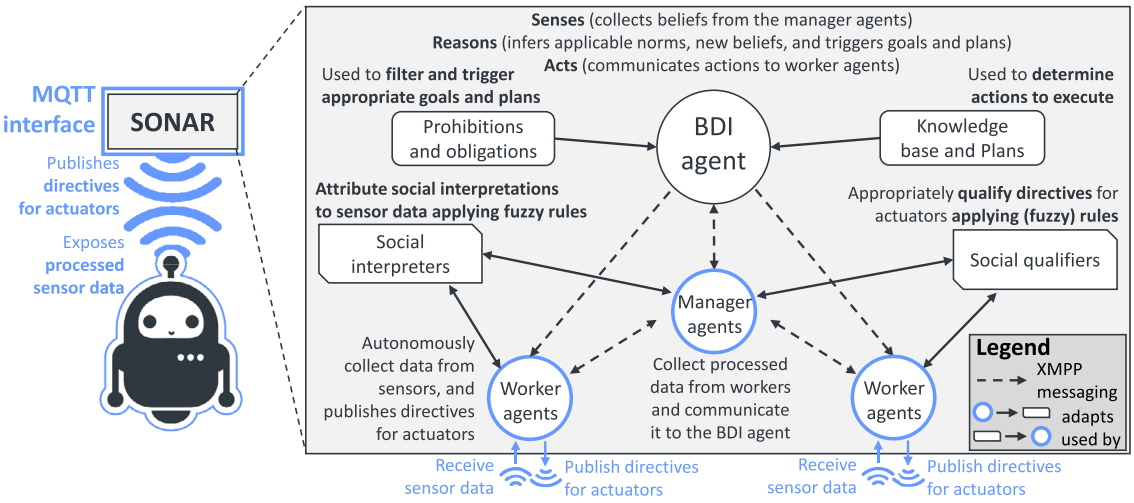


Fig. 2 Illustration of SONAR, the adaptive control architecture for SOcial Norm Aware Robots

[10] describe an approach, based on the concept of institutions, to introduce normative aspects into robot behaviors for mixed human-robot societies that adhere to human-defined norms. Fuzzy logic has recently shown potential for normative modeling and reasoning [12, 13, 25]. Bruno et al. [11], propose a framework for culture-aware robots based on fuzzy logic control. Similarly, Vitiello et al. [26] have shown that fuzzy logic and fuzzy inference can be effectively used by social robots to autonomously reason about, and react to, social norms such as proxemics behaviors, i.e., to determine the appropriate distance to keep from humans in different circumstances based on cultural or individual preferences. Besides some primarily theoretical attempts for bringing BDI and fuzzy logic together [73–81], little or no work exists on the combination of BDI and fuzzy logic in the context of norm-aware social robots.

3 SONAR: Proposed Architecture for Making Social Robots Norm-Aware

This section describes SONAR, the proposed adaptive control architecture for social norm-aware robots. First, we explain the main elements of SONAR, and how these elements interact with each other. Next, we provide the details on how the architecture allows for social and norm-aware reasoning. Finally, we discuss how learning and adaptation of the norms may occur in SONAR.

3.1 Main Elements of SONAR

Figure 2 illustrates the main elements of the architecture. SONAR follows the design principles for the development of intelligent rational cognitive social agents identified by Rato et al. [43] and Castro et al. [46], summarized in Sect. 2. In SONAR, first the inputs perceived by the robot via its sensors from the environment are transformed into *beliefs* that characterize the current operating context (see worker and manager agents in Fig. 2), and are given a social interpretation using fuzzy rules (see social interpreters in Fig. 2). These rules characterize the social norms for the interpretation of the physical reality [36, 82]. Moreover, before execution, the actions of the robot are assessed through a social qualification procedure in order to ensure that they are socially *adequate* based on the identified social context (see social qualifiers in Fig. 2).

Technically speaking, SONAR is a Multi Agent System (MAS) [83], where multiple agents *autonomously* and asynchronously operate and interact with each other via message passing.[2] Designing SONAR as a multi-agent system ensures a distributed execution of the different components. Besides extensibility, maintainability, and flexibility, this also implies computational efficiency. Different agents within SONAR can technically run on entirely different machines, including dedicated high-performing clusters, if needed. Three types of agents operate in SONAR according to their tasks and roles that characterize a hierarchy in the MAS: the *worker agent* type, the *manager agent* type (a special type of worker agent) and the *BDI agent*. Figure 2 depicts two worker agents, one manager agent and one BDI agent. The number of worker and manager agents is meant for illustrative purposes and aims at clearly showing the hierarchy of agents in the MAS. However, SONAR does not pose any restriction on the number of worker and manager agents. While it is technically possible for multiple BDI agents to exist in SONAR if adequately coordinated, in this paper we consider only one BDI agent

[2] Our python implementation of SONAR makes use of the SPADE multi-agent systems platform [84], where communication between agents is based on instant messaging (XMPP [85]) which supports the FIPA Agent Communication specification [86] metadata.

that handles the main reasoning cycle of the robot. Next, we explain these different types of agents.

Worker agents Worker agents are MQTT[3] clients. Worker agents regularly and autonomously collect and publish data from and to the MQTT broker. Worker agents subscribe to MQTT topics to receive sensor data exposed by an MQTT broker, and they send directives to the robot actuators by publishing such directives to the MQTT broker. Worker agents can process, aggregate, and modify the data according to their particular tasks (e.g., a *chatter agent* deals with communication-related tasks, while a *norm adapter agent* deals with norm adaptation). Additionally, each worker agent cyclically performs a default behavior wherein it asynchronously awaits messages from other agents without blocking the execution of its own tasks or those of other agents.[4] Upon receiving a message, the agent processes it accordingly. For example, a *posture handler agent* awaits directives and information from the *BDI agent* on adjusting some of the robot's actuators, such as rotating its head. A *vision handler agent* awaits requests from a *manager agent* to communicate the most recent vision-related data, such as detected objects.

Manager agents Manager agents regularly–i.e, at periodic intervals– request information from relevant worker agents, which are specified for each manager agent, and represent their data sources. The length of the interval depends on the types of data collected by the manager agent from the worker agents, and on the need to ensure adequate real-time human-robot interactions.[5] After requesting data from their data sources, the manager agents shortly (at most until the end of the current interval) and asynchronously wait for data to be received. Once data is received by all data sources (or once the timeout is reached), the manager agents aggregate the data and produce beliefs to communicate, when requested, to the BDI agent. Since a manager agent is a special type of a worker agent, it also cyclically awaits messages from other agents. In particular, manager agents await a request for new beliefs from the BDI agent. Manager agents have more direct communications with the BDI agent than worker agents. This helps minimizing the communications between the BDI agent and other agents, and allows manager agents to prepare data for the BDI agents that requires information from multiple sources. For example, in order to create a message `said(davide, hello)`, it is necessary to collect data from both the camera of the robot (for detecting the face of the human and identifying their identity, in this case Davide) and from the microphones (for identifying the verbal message "hello" communicated via the human).

BDI agent The BDI agent cyclically performs the *sense*, *reason*, and *act* deliberation activities [47]. During the *sense* activity, the agent requests to the (manager) agents to send new *perceived beliefs*, i.e., beliefs that are inferred based on the data perceived via the sensors of the robot. The set of data that is perceived at a certain instant via all sensors of the robot is referred to as *context*, because, from the robot's perspective, such data characterizes the circumstances in which the robot is operating. If new perceived beliefs are communicated, the BDI agent performs the *reason* activity: First, (i) the BDI agent uses the perceived beliefs in order to infer, via belief propagation rules, additional beliefs that can be inferred (e.g., if the perceived belief is that person p is visible, then the agent can infer the belief that p is the person the robot should interact with). Every time a (perceived or inferred) belief is generated, the BDI agent stores it both in its belief base and in a short-term memory module[6]. In the belief base, this belief is used for reasoning and is revised when new beliefs are generated. The short-term memory module tracks the previous beliefs and observations (e.g., to spontaneously trigger a conversation about an object that has been perceived for the first time in recent memory). Then, (ii) the BDI agent performs social and normative reasoning via inference rules that determine which norms apply in the current context, which actions and goals are prohibited or obliged, and what the social role of the robot is. Finally, (iii) the BDI agent triggers goals, and selects plans according to its plan base and to the active norms. The execution order of the concurrent plans, and the mechanisms to handle conflicting information, both depend on the design of the BDI agent. For instance, the BDI agent may be designed to set plan priorities through the rule ordering in AgentSpeak [53].[7] In our implementation of SONAR evaluated in the experiments described in Sect. 4, the priority given to different aspects during reasoning is as follows: `greeting` $\succ$ `robot commands` (e.g., to shut down) $\succ$ `posture` $\succ$ `perceived interlocutor interest` (e.g., inferred from the gaze) $\succ$ `perceived objects` $\succ$ `developing trust` $\succ$ `proactive speech` $\succ$ `reactive speech`. During the *act* activity, the agent executes the actions that are in line with the intentions inferred via the *reason* activity. This is done by composing the plans that are chosen from the plan base, such that the current goals

[3] MQTT is an ISO recommended [87] lightweight machine-to-machine network protocol, particularly used in the context of internet of things (IoT). By adopting this type of protocol, we fully decouple SONAR from the specifics of the robot, therefore making SONAR a robot-agnostic architecture.

[4] In our SPADE-based implementation of SONAR, this is achieved via async coroutines [84].

[5] In our implementation 0.2 s.

[6] In our implementation, the short-term memory module is realized via a dictionary of key-value pairs, with timestamps as keys and beliefs as values. To preserve the computational efficiency and to minimize the computational overhead, beliefs older than one minute (an adjustable parameter of the BDI agent) are regularly deleted.

[7] The language used to encode the plan base of BDI agents in our SPADE-based implementation of SONAR [84].

are achieved. If the actions composing a plan need to be performed by worker agents (e.g., because they involve the use of the actuators of the robot), then the BDI agent communicates the actions to worker agents.

3.2 Social and Normative Reasoning

SONAR supports the following three types of rules that are used to model (social) norms and to perform social and normative reasoning: *social interpretation rules*, *behavior qualification rules*, and *prohibition and obligation rules*. Next, we explain these three categories of rules in SONAR.

Social interpretation rules SONAR uses *social interpretation rules* to associate the social and situational cues (e.g., the distance between people and/or agents during a conversation) with social meanings (e.g., the DIAMONDS situation characteristics given in [69]). These associations are fuzzy in their nature (e.g., different values for the distance can be considered as *low* for different people), and they might differ from a context or a culture to another [21]. Therefore, to represent these associations we use IF-THEN fuzzy rules of the form "IF c_1 AND ... AND c_q, THEN m_1 AND ... AND m_k", with $c_1, \ldots, c_q$ and $m_1, \ldots, m_k$ generally given by the formulation "a IS b", which contains linguistic terms. More specifically, such a formulation indicates that a *linguistic/qualified value*, b, is assigned to a *linguistic variable*, a. An example of such a fuzzy rule is "IF *distance* IS *Low*, THEN *positivity* IS *High-positive-correlation*", where *distance* and *positivity* are linguistic variables representing, respectively, a social cue and a situation characteristic, and *Low* and *High-positive-correlation* are linguistic values for those variables. Intuitively, the example indicates that maintaining a close proximity (*low distance*) during a conversation can be interpreted, socially, as strongly indicating (*high-positive-correlation*) a *positivity*-related ([69]) situation. Mathematical realizations of linguistic values in fuzzy logic are *fuzzy sets* that are represented via *membership functions*[8]. Membership functions specify the degree (called *degree of truth*) to which a crisp measurement of a *base* variable (e.g., 2 meters for base variable Distance) is member of a particular fuzzy set that represents a linguistic term. For instance, 2 meters is *low* with a degree of truth of 0.8, and is *medium* with a degree of truth of 0.2. Membership functions, therefore, allow to quantify approximate linguistic terms, and they can be defined by a system designer (e.g., based on existing knowledge about that particular linguistic concept), or may be learnt over time and in the course of using the fuzzy rule base in various interactions (as we will discuss later in this paper).

In SONAR, the set of input data D_t (e.g., the measured distance between the robot and a human, a detected sound, or the speech decoded from the sound) received at time instant t by the manager agent is used to determine, via fuzzy inference[9], a set O_t of fuzzy membership degrees $\mu(S_i)$ for all social interpretations S_i for $i = 1, \ldots, \rho$, where the number ρ is the number of possible social interpretations of a situation (e.g., $\rho = 8$ if the 8 DIAMONDS are considered based on [69]). For instance, given a measured *distance* of 2 meters and a value 1 for a binary variable *communicating* (indicating that the situation involves communication), it is inferred that the situation can be interpreted as related to *sociality* with degree of truth 0.8, to *positivity* with degree of truth 0.6, to *negativity* with degree of truth 0.2, etc. The set O_t, therefore, contains information about the degree of truth of possible social interpretations of a situation. This set can directly be used in normative reasoning and decision-making via SONAR, e.g., as input for performing fuzzy inference via the behavior qualification rules to determine, via defuzzification, adequate parameters for the robot's actuators.

Behavior qualification rules A robot that is placed in a social context is not only expected to give an appropriate social meaning to physical inputs, but also to act in a way that is considered socially acceptable and in line with social norms and practices. We represent the behavior qualification rules via a combination of fuzzy and non-fuzzy rules, and use them to determine appropriate (norm-aligned) *qualifiers* of behavior (i.e., directives for the actuators of the robot). For example, a chatter agent (which is a specific type of worker agents explained in Sect. 3.1) that has been instructed to convey a message via chatting to the human, will send a directive to the robot interface that includes a sentence, as well as the qualifiers (e.g., the adequate volume of the voice, the pitch, the speed of talking) that are inferred as appropriate in the current situation (e.g., are interpreted as *Social*), using the behavior qualification rules.

Prohibition and obligation rules Prohibition and obligation rules are given as tuples $\langle s_n, z_n, t_n \rangle$ for $n \in N$, with N the set of all norms, s_n a conjunction of *beliefs* that characterizes the conditions for applicability of norm n, $z_n \in \{oblig, prohib\}$ indicating whether or not norm n is an obligation or a pro-

[8] A membership function is defined by $\mu : U \rightarrow [0, 1]$, where U is the *universe of discourse* (i.e., the range of all possible crisp values) for the linguistic variable. For instance, $U = [0, 10]$ (in meters) can be the universe of discourse for fuzzy variable *Distance*.

[9] For the sake of brevity, we omit an in-depth discussion about fuzzy inference systems (FISs). Briefly speaking, given the crisp values of input variables, a FIS first applies a procedure called fuzzification on the input data in order to transform these crisp values into fuzzy values. Next, relevant fuzzy operations are performed on the fuzzy rules of the rule base that have been fired by the input data, in order to make an inference and determine the output in terms of fuzzy sets. These fuzzy outputs are then converted into crisp values, which is called the defuzzification procedure. In our implementation of SONAR, we make use of the Mamdani max-min inference method. More details about FISs can be found in [23].

hibition (where *oblig* and *prohib* are two labels representing an obligation and a prohibition, respectively), and t_n is the target of the norm, i.e., either an *action* (part of a plan) or a *goal* of the BDI agent. An example of obligation targeting a goal is $\langle social\_distance \wedge socially\_related\_situation \wedge not\_greeted(person), oblig, greet(person)\rangle$. This obligation represents the norm for greeting behavior, i.e.: "*It is appropriate to greet whenever a person is visible at a social distance, the situation is considered socially-related, and no greetings has occurred yet.*". An example of related prohibition targeting an action, instead, is $\langle conversation\_start \wedge not\_greeted(person), prohib, update\_topic\rangle$. This prohibition represents the dialogue norm "*At the beginning of a conversation, it is not appropriate to start talking about any topic before greeting.*". Prohibition and obligation rules are used both to trigger new goals and plans (therefore making a robot proactive), and to select appropriate goals and plans (to make the robot reactive) during the norm-aware reasoning of the BDI agent. For instance, the norm-aware reasoning of the BDI agent concerning the example of obligation above can be represented via the following AgentSpeak rule:

```
+!reason_about_greeting_norms:
    visible(face, Person) &
    distance(Person, social)
<- !greet(Person).
```

The rule indicates that whenever the agent has the goal `!reason_about_greeting_norms` to reason about greeting norms (a goal that is created by the BDI agent during the *reason* deliberation activity explained in Sect. 3.1), and believes that the face of a person is visible at a social distance, then a new goal `!greet(Person)` is created. Then, during the *act* activity, the agent will attempt to achieve the goal by means of a plan, e.g., by means of an action `.greet(Person)` that will instruct the chatter worker agent to begin a greeting procedure.

3.3 Learning and Adaptation to Norms

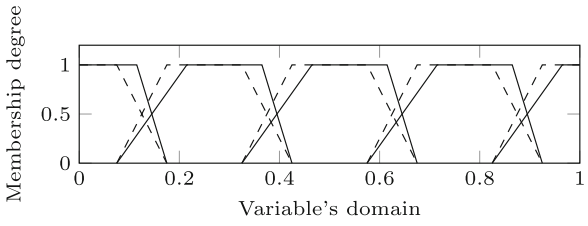
In this section, we introduce the norm adaptation mechanisms that are supported by SONAR. In Sect. 6, we will illustrate via computer-based simulations that a robot endowed with the proposed mechanisms can quickly adapt to the norms of a society. We focus on adaptation of the norms that are represented via fuzzy rules, i.e., the social interpretation rules and the fuzzy behavior qualification rules explained in Sect. 3.2. More specifically, we focus on adaptation of the linguistic variables that compose the fuzzy rules. We do not focus instead on learning new fuzzy rules nor on adaptation and learning strategies for prohibitions and obligations, for which some solutions can be found in the literature (e.g., [25, 27, 88]).

Norm adaptation in SONAR is performed by a *norm-adapter* agent. The norm-adapter agent is a type of a worker agent that adjusts the membership functions, which mathematically represent the fuzzy sets that model the linguistic values corresponding to the norms. This norm adaptation is based on the data that has been collected throughout the human-robot interactions, or via observations of human-human interactions. Every time a dataset D_t is collected by a manager worker for time instant t (resulting, via the application of social interpretation rules, in the set O_t of degrees of truth of possible social interpretations of the situation), the social interpretation $S_t^\star$ for time instant t with the highest degree of truth is determined: $S_t^\star = S_i$, where $\mu(S_i) > \mu(S_j)$ for all $j \neq i$, and for $i, j = 1, \ldots, \rho$, and with $\mu(S_i), \mu(S_j) \in O_t$, randomly selecting one of the equally true interpretations in case of ties. The social interpretation $S_t^\star$ is then communicated to the norm-adapter agent together with the dataset D_t. The norm-adapter agent regularly–at periodic intervals– examines the collected data and initiates a norm adaptation algorithm once a sufficient amount (that is pre-defined) of data has been collected. Every time the adaptation process concludes, the updated membership functions are made available to the other worker agents, by updating the social interpreters and the social qualifiers in Fig. 2. In the following, we explain in details how the norm adaptation works.

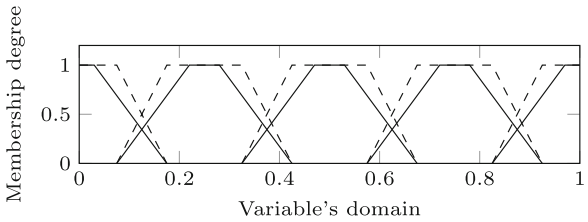
3.3.1 Norm Adaptation via Fuzzy Sets Modification

Given a fuzzy rule that indicates "IF *sociality* IS *High-positive-correlation*, THEN *distance* is *Medium*", our goal is to learn the (membership function of the) fuzzy set that represents the concept of *Medium* for distance, based on the data that is collected by the robot via observing, or interacting with, the individuals from a society.

We call the variables that are subject to adaptation *dynamic (linguistic) variables*, where these variables characterize the subjective, personal, or cultural aspects of the fuzzy rules. We represent the fuzzy sets corresponding to the dynamic variables via trapezoidal membership functions, which are defined by four parameters s_l, c_l, c_u, s_u, with $s_\mathrm{l} \leq c_\mathrm{l} \leq c_\mathrm{u} \leq s_\mathrm{u}$, where s_l and s_u are, respectively, the lower and upper bounds of the support (i.e. the base), and c_l and c_u are, respectively, the lower and upper bounds of the core of the trapezoidal functions. A *partition* P_v for the linguistic variable v is the set $P_v = \{F_1, \ldots, F_p\}$ of the p fuzzy sets (linguistic values) F_i for $i = 1, \ldots, p$ that characterizes the domain of the linguistic variable. For instance, a partition for the variable *distance* defined on a given domain (e.g., [0, 10] meters) may be composed of three fuzzy sets *Low*, *Medium*, and *High*, for which the corresponding membership functions cover the given domain.



(a) Core-position modifier, $k_{\mathrm{CP}} = 0.8$



(b) Core-width modifier, $k_{\mathrm{CW}} = -0.6$

Fig. 3 Examples from [89] for modification of a partition that is composed of five trapezoidal fuzzy sets. The dashed plots illustrate the initial membership functions, whereas the solid plots correspond to the modified membership functions

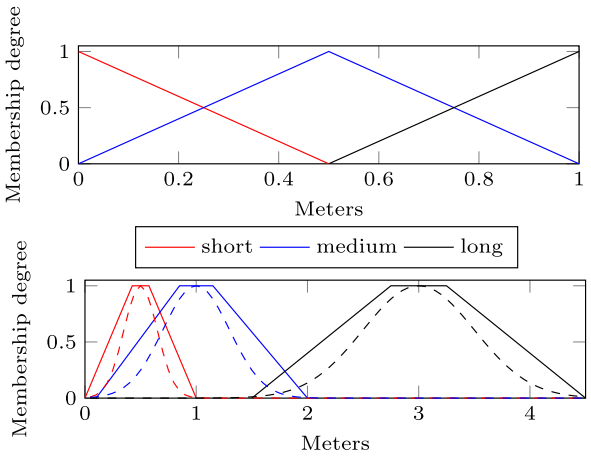


Fig. 4 Initial (before norm adaptation) membership functions for dynamic variable *distance* that need to be adapted (top plot), and a desired outcome after execution of the norm adaptation (bottom plot): Solid curves represent the estimated functions, whereas dashed curves show the fuzzy Gaussian membership functions that represent the real distributions of the data points that are collected for the training/adaptation procedure

Our approach for adaptation to the norms is a modification of the context adaptation technique introduced by Botta et al. [89]. Figure 3 illustrates two examples for adaptation of the position of the core and width of the trapezoidal membership functions, where these functions may correspond to the fuzzy sets that represent the linguistic terms that are used in the rules that incorporate the norms. In SONAR, the norm-adapter worker agent modifies the membership functions attempting to reduce the error that is resulting from using these membership functions, compared to the data collected by the robot. The top plot in Fig. 4 illustrates the initial representation for the membership functions of the fuzzy sets within the partition of all those dynamic variables for which no training data is available yet. Note that all membership functions are defined as trapezoidal functions within the domain [0, 1]. The bottom plot in Fig. 4 shows the adapted membership function (see the dashed curves) for a particular dynamic variable when the collected data has been used to train the membership functions. In SONAR, we consider the *ideal adaptation* to be such that the center and the width of the core for the i-th trapezoidal function in partition of a dynamic variable correspond to, respectively, the mean and the standard deviation of available data about the corresponding linguistic term (e.g., about *Medium* distances), and that the domain of the trapezoidal function includes all the corresponding observed values. In Fig. 4, the domain of the adapted functions (bottom plot) is different from that of the initial functions (top plot). This illustrates that the adaptation mechanism is independent from the domain of the variables, and is made possible by scaling the functions according to the observed data.

Finally, although we use trapezoidal membership functions, we remark that our approach, in line with the work from Botta et al. [89] that we use as a starting point, can easily be adapted to work with any other shapes of membership functions. Given that our study primarily focuses on the exploratory aspect of modifying membership functions for the aim of norm adaptation, and does not focus on a particular domain, we choose trapezoidal fuzzy sets [90] to provide a generalized solution that accommodates various types of membership functions, since other commonly used membership functions, such as triangular and singleton functions, are special cases of trapezoidal functions. The rationale of our adaptation approach remains the same also for gaussian shaped membership functions.

3.3.2 The Norm Adaptation Algorithm

The norm-adapter agent regularly examines the data set D_S collected per social interpretation S (e.g., *duty* or *sociality*). When the number of the data points in the data set D_S reaches a given threshold (say τ), then the agent enacts a norm adaptation algorithm for interpretation S, by executing the following steps. An illustrative example of execution of the algorithm is reported in Fig. 5.

Step 1. Determine the set R_S of fuzzy rules that are related to social interpretation S. For instance, if S is *Sociality*, rules in R_S contain, either in the premise or in the consequent of the rule, an assignment that characterizes a positive correlation with S (e.g., *Sociality* IS *High-positive-correlation*), and an assignment for at least one dynamic variable (e.g., *Distance*

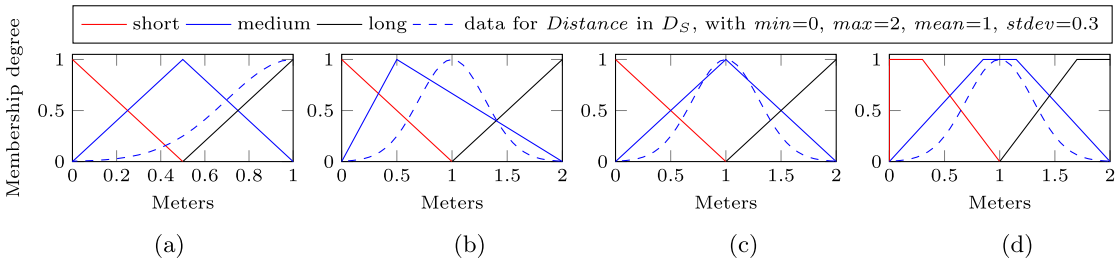


Fig. 5 Example of the execution of the algorithm for norm adaptation given a dataset D_S, with $|D_S| \geq \tau$, collected for social interpretation $S = \textit{Sociality}$. For this example, *Step 1* determines the set $R_S = \{r\}$ containing only one rule $r =$ IF *Sociality* IS *High-positive-correlation* THEN *Distance* IS *medium*. *Distance* is a dynamic variable that needs to be adapted. Its partition is composed of the three membership functions in Fig. 5a (solid curves). *Step 2* determines $V_r = \{\textit{Distance}\}$, $\textit{Distance}_L = \textit{medium}$, and $c_{\mathrm{u}}^{\textit{Distance}_L} = c_{\mathrm{l}}^{\textit{Distance}_L} = 0.5$. *Step 3* (shown in Fig. 5b) scales linearly the universe of discourse of *Distance* (compare the domain of the functions between Fig. 5a and b), and the support of all the membership functions in its partition according to the data collected for *Distance* (represented via the blue dashed curve). *Step 4* (Fig. 5c) modifies the position of the core of all the membership functions based on the error between the current center of the *medium* fuzzy set (referred in rule r), and the mean *Distance* in data, resulting in $k_{\mathrm{CP}} = 0$ for all fuzzy sets, and a right shift of 0.5 of the center of the *medium* fuzzy set. *Step 5* (Fig. 5d) modifies the width of the core of all the membership functions based on the error between the current width of the *medium* fuzzy set and the standard deviation of *Distance* in data, resulting in $k_{\mathrm{CW}} = 0.3$ for all fuzzy sets, and a dilation of the core width of all the fuzzy sets

IS *medium*). Then for each rule $r \in R_S$ execute the following steps.

Step 2. For all dynamic variables v within the set V_r of dynamic variables that appear in rule r perform the adaptation procedure (i.e., go to *Step 3*), unless adaptation for the same dynamic variable has already been performed via another rule. In the following, for all dynamic variables v and rule r, we call v_L the fuzzy set of the variable v referred in rule r (e.g. $v_L = \textit{medium}$ for $v = \textit{Distance}$ if the rule contains *Distance* IS *medium*),[10] and $c_{\mathrm{u}}^{v_L}$ and $c_{\mathrm{l}}^{v_L}$, respectively, the lower and upper bounds of the core of v_L.

Step 3. Scale linearly the universe of discourse of variable $v \in V_r$ and all the supports of the membership functions corresponding to the partition of variable v, so that to reflect the boundaries of the measurements that have been collected for that variable. We use the following standard linear scaling function:

$$s : [a, b] \rightarrow [a', b']$$
$$s(v) = a' + (b' - a') \cdot \frac{v - a}{b - a}, \quad \forall v \in [a, b]$$

where parameters a and b identify the bounds of the original universe of discourse, and a' and b' identify the bounds of the new universe of discourse obtained from the new measurements. We compute the new boundaries for $[a', b']$ via

$$a' = \min\{a, v_{\min, D_S}\}, \quad b' = \max\{b, v_{\max, D_S}\}$$

where $v_{\min, D_S}$ and $v_{\max, D_S}$ are, respectively, the minimum and maximum values of variable v observed in data set D_S.

[10] We assume that in every fuzzy rule, at most one linguistic value is assigned to each linguistic variable.

Step 4. Modify the position of the core for all membership functions corresponding to the partition of dynamic variable v by shifting the core within the support while maintaining the original width (i.e., the distance between c_{l} and c_{u}) using the following relationship:

$$c'_{\mathrm{l}} = \begin{cases} c_{\mathrm{l}} - (s_{\mathrm{l}} - c_{\mathrm{l}}) \cdot k_{\mathrm{CP}} & \text{if } k_{\mathrm{CP}} < 0 \\ c_{\mathrm{l}} + (s_{\mathrm{u}} - c_{\mathrm{u}}) \cdot k_{\mathrm{CP}} & \text{if } k_{\mathrm{CP}} \geq 0 \end{cases}$$
$$c'_{\mathrm{u}} = \begin{cases} c_{\mathrm{u}} - (s_{\mathrm{l}} - c_{\mathrm{l}}) \cdot k_{\mathrm{CP}} & \text{if } k_{\mathrm{CP}} < 0 \\ c_{\mathrm{u}} + (s_{\mathrm{u}} - c_{\mathrm{u}}) \cdot k_{\mathrm{CP}} & \text{if } k_{\mathrm{CP}} \geq 0 \end{cases}$$

where c'_{l} and c'_{u} are, respectively, the lower and upper bounds of the modified core, and $k_{\mathrm{CP}} \in [-1, 1]$ is a parameter that characterizes the intensity of the shift for the lower bound (whenever $k_{\mathrm{CP}} < 0$) or for the upper bound (whenever $k_{\mathrm{CP}} > 0$). We define the *core-position error* ϵ_{CP} as the difference between the current center of the core of the membership function of v_L and the mean v_{mean, D_S} of the values observed for variable v within data set D_S, i.e., $\epsilon_{\mathrm{CP}} = \frac{c_{\mathrm{u}}^{v_L} - c_{\mathrm{l}}^{v_L}}{2} - v_{\mathrm{mean}, D_S}$. We determine k_{CP} as the inverse of the core-position error *ratio*. More specifically,

$$k_{\mathrm{CP}} = \begin{cases} -1 \cdot \min(1, \frac{\epsilon_{\mathrm{CP}}}{c_{\mathrm{l}} - s_{\mathrm{l}}}) & \text{if } \epsilon_{\mathrm{CP}} \geq 0 \\ -1 \cdot \max(-1, \frac{\epsilon_{\mathrm{CP}}}{s_{\mathrm{u}} - c_{\mathrm{u}}}) & \text{if } \epsilon_{\mathrm{CP}} < 0 \end{cases}$$

Step 5. Modify the width of the core for all membership functions corresponding to the partition of dynamic variable v by dilating or shrinking the core of the membership function within the support using the following relationship:

$$c'_{\mathrm{l}} = \begin{cases} c_{\mathrm{l}} + w \cdot (s_{\mathrm{l}} - c_{\mathrm{l}}) \cdot k_{\mathrm{CW}} & \text{if } k_{\mathrm{CW}} < 0 \\ c_{\mathrm{l}} + (s_{\mathrm{l}} - c_{\mathrm{l}}) \cdot k_{\mathrm{CP}} & \text{if } k_{\mathrm{CW}} \geq 0 \end{cases}$$

$$c'_{\mathrm{u}} = \begin{cases} c_{\mathrm{u}} + w \cdot (s_{\mathrm{u}} - c_{\mathrm{u}}) \cdot k_{\mathrm{CW}} & \text{if } k_{\mathrm{CW}} < 0 \\ c_{\mathrm{u}} + (s_{\mathrm{u}} - c_{\mathrm{u}}) \cdot k_{\mathrm{CW}} & \text{if } k_{\mathrm{CW}} \geq 0 \end{cases}$$

where $w = (c_{\mathrm{u}} - c_{\mathrm{l}})/(c_{\mathrm{l}} - s_{\mathrm{l}} + s_{\mathrm{u}} - c_{\mathrm{u}})$, and $k_{\mathrm{CW}} \in [-1, 1]$ is a parameter that characterizes the intensity of the dilation (whenever $k_{\mathrm{CW}} > 0$) or the shrinkage (whenever $k_{\mathrm{CW}} < 0$) of the core. We define the *core-width error* ϵ_{CW} as the difference between the current width of the core of the membership function of v_L and the standard deviation v_{sd,D_S} of all the values observed for dynamic variable v within data set D_S, i.e., $\epsilon_{\mathrm{CW}} = (c_{\mathrm{u}}^{v_L} - c_{\mathrm{l}}^{v_L}) - v_{\mathrm{sd},D_S}$. We determine k_{CW} as the inverse of the core-width error *ratio*. More specifically,

$$k_{\mathrm{CW}} = \begin{cases} -1 \cdot \min(1, \frac{\epsilon_{\mathrm{CW}}}{w \cdot (c_{\mathrm{l}} - s_{\mathrm{l}})}, \frac{\epsilon_{\mathrm{CW}}}{w \cdot (s_{\mathrm{u}} - c_{\mathrm{u}})}) & \text{if } \epsilon_{\mathrm{CW}} \geq 0 \\ -1 \cdot \max(-1, \frac{\epsilon_{\mathrm{CW}}}{c_{\mathrm{l}} - s_{\mathrm{l}}}, \frac{\epsilon_{\mathrm{CW}}}{s_{\mathrm{u}} - c_{\mathrm{u}}}) & \text{if } \epsilon_{\mathrm{CW}} < 0 \end{cases}$$

Finally, when the membership functions for all the fuzzy sets within the partition of dynamic variable v are modified by the norm-adapter agent, these are made available to the other worker agents, who use them for social interpretation and behavior qualification.

4 Case Study: Interaction of Humans with a SONAR-Based NAO Robot

In this section, we discuss our extensive exploratory case study that was designed to demonstrate the feasibility and applicability of SONAR for human-robot interactions. We do so, by assessing the *effectiveness* and *efficiency* [91] of our Python 3.9 implementation of SONAR, and the *perception, experience*, and *acceptance* of the robot (which is steered via such implementation of SONAR) by the participants of the experiments. In this set of experiments, we excluded the norm adaptation procedure from SONAR, for the following two reasons: First, to focus on only the SONAR architecture independently, without integrating it with an adaptation algorithm. Second, adaptation of the fuzzy membership functions requires enough data and thus multiple interactions with each participant. Since in our setup, we were not able to recruit the participants for more than one session, such long-term interactions had to happen in only one session. This makes it likely that participants get exhausted, which may falsely affect the criteria of assessment. In real-life applications, a companion robot for instance, will spend more time with its users. Thus gathering the data that is required for adaptation of the fuzzy membership functions will not result in such issues. Therefore, we evaluate the norm adaptation procedure separately in Sect. 6 via extensive computer-based simulations.

In this case study, we address the following research questions:

RQ1.1: *To what extent is SONAR usable for the real-time control of a social robot that accounts for situation cues and norms during interactions with humans?*

RQ1.2: *What is the human perception, experience, and acceptance of a social robot that employs SONAR with the aim of considering situation cues and norms in its decision-making and exhibiting proactive behaviors?*

To investigate these research questions, we conducted an experiment where adults interacted with a Nao robot [33] in a conversation scenario. Two contrasting behavior styles for the robot were considered, which we refer to as Nao-Chatbot and Nao-SONAR (details are given below). We collected both quantitative and qualitative feedback from the execution logs that were generated by the robot during the experiments and via questionnaires that were completed by the participants before and after interacting with the robots.

4.1 Methodology of the Case Study

Next, we explain our methodologies for designing and executing the experiments.

4.1.1 Human Participants

Individual human participants took part in this study during December 2022. The experiments took place in 4 meeting rooms of the Faculty of Aerospace Engineering of TU Delft. A commercially available humanoid Nao robot v6 [33] was used for the experiments. The participants had an open-ended conversation with the robot within the context of five specific tasks (see the *Main Trial Phase* in Sect. 4.1.2). Figure 1 illustrates the setup: For each experiment, one participant was seated in front of Nao, which was standing on a table. On the table, four objects were placed: a captain hat, a plant, a bottle, and a teddy bear. The meeting rooms also had a monitor and a clock (not visible in Fig. 1) placed on the wall.

In total, a sample of 25 adult volunteers (52% female, 48% male) was recruited from the Delft University of Technology. The age of the participants ranged between 18 and 64 (with 8% between 18 and 24, 72% between 25 and 34, and 8% between 55 and 64). Their education level ranged from high school diploma to doctorate (with 8% high school graduate, 8% BSc degree, 72% MSc degree, and 12% doctorate). From the participants, 16% were university support staff, 8% were students, 68% were PhD students or researchers, and 8% were academic or faculty staff. The self-reported information about the familiarity of the participants with robots before the experiment included: 32% not familiar at all, 40% slightly familiar, 16% moderately familiar, 12% very familiar, and 0% extremely familiar. All participants completed the consent forms that are provided as an attachment to this paper. One participant did not agree to record the video of the

interactions with the robot. The participants were not paid for their participation in the experiments.

4.1.2 Experimental Procedure

Each experiment was composed of the following 3 phases: introduction phase, main trial phase, and final phase. These are explained in detail below.

Introduction phase Before the start of each experiment, a general introduction phase was performed, where the robot was presented to the participant, showcasing some of its basic movements and general capabilities, so that the participant got acquainted with the robot before the start of the experiment. An information sheet was given to each participant to read, in order to understand the basic principles of the interaction with the robot, along with a consent form to be signed. After signing the consent form, the participant was requested to complete an Introductory Questionnaire and a NARS (Negative Attitude towards Robots Scale) Questionnaire [92, 93], which are briefly described below. All questionnaires and information sheets provided to the participants are anonymized and made available in our online appendix [34].

Introductory Questionnaire. This questionnaire includes 7 questions for collecting information about the gender, age range, occupation, education, level of familiarity with robots (using a Likert scale), prior experience with companion robots (e.g., at work, as toys, via movies, books, or TV shows, in museums or at school, in person), and level of technical knowledge with robots.

NARS (Negative Attitude towards Robots Scale) Questionnaire. This questionnaire includes 16 questions for measuring the attitudes of humans towards robots in daily life. The answers to this questionnaire are used to highlight any potential prior (negative) bias of the participants in their attitude towards robots. The results of this questionnaire were used in our experiments to validate the randomization of the experiments.

Main trial phase The main trial phase consisted of an open-ended conversation with the robot. Additionally, the participants were instructed to perform the following 5 specific tasks (see Table 1 for more details) during their conversation with the robot: greeting, role playing game, discussing a personal issue, paying attention to an object, goodbye. These five tasks aimed to assess the effectiveness of the robot in adapting to different situations, by leveraging its awareness of social rules and environmental cues. Each task also provided an opportunity to assess various technical aspects of our implementation related to social and norm awareness (see Table 1, last column), and to the behavioral requirements for social and norm-aware robots, as highlighted in Sect. 2.

Tasks 1 and 5 (greeting and goodbye) focused on standard moments of a casual conversation and served to define clear experimental boundaries for participant interactions. The participants had full control over the Main trial phase completion, without the experimenter being present in the room.

Task 2 (role awareness) exemplified the societal notion that specific responsibilities and behaviors are dictated by social roles [94]. In fact, a social and norm-aware robot is expected to adapt its behavior according to the role of its interlocutor [43].

Task 3 (trust) underscored the importance of social robots being able to establish adequate trust in interactions with humans [9, 95–98], which can be facilitated by norm compliance [30, 99].

Finally, Task 4 (social cues and environment awareness) addressed the necessity for social robots to interpret implicit or explicit social cues that are provided by humans and to reason about these cues within the context of their environment, in order to ensure natural and meaningful interactions with humans [43, 64].

For every participant, the order of tasks 2–4 was randomized in order to test SONAR on a variety of combinations of behaviors. After performing the tasks, the participant was asked to complete two questionnaires based on the COGNIRON Robot Personality Questionnaire [93] and the USUS framework [100], which are explained below.

Extended COGNIRON robot personality questionnaire. This questionnaire is used to evaluate the attribution of each of the following personality characteristics to a robot, using a 5-point Likert scale: anxiety, tension, shyness, vulnerability, sociability, general activity level, assertiveness, excitement seeking, dominance, aggressiveness, impulsiveness, creativity, autonomy, intentionality, predictability of behavior, controllability, and considerateness. We extended the original questionnaire with 3 additional questions concerning the reactiveness, proactiveness, and autonomy of the robot, in order to assess the major aspects that traditionally characterize intelligent agents within the AI literature [83].

USUS-Based questionnaire. This questionnaire is composed of 45 questions (that should be answered using a 5-point Likert scale) and is designed based on the USUS (Usability, Social Acceptance, User Experience, Societal Impact) framework [100]. We tailored this questionnaire for our particular case study, considering the first three aspects of the USUS framework, where the latter (i.e., Societal Impact) was assessed at the end of the entire experiment, as part of the Final Questionnaire explained later on in this section.

The Main Trial Phase was repeated twice per participant, considering Nao-Chatbot and Nao-SONAR as the behavior styles of the robot. The order of the exposure of the participants to the robots with these two behavior styles was randomly determined per participant, where 52% of the par-

Table 1 Specific tasks considered for the human-robot interactive conversation, including instructions for the participants, as well as the expected behavior for Nao-SONAR (last column)

Task	Instructions for the participant	Expected behavior for Nao-SONAR (in case of success with respect to the desired metrics)
Task 1 (greeting)	Task 1: Act similarly to when you are in a normal greeting situation with somebody (in this case the Nao robot) you meet for the first time	(i) Establish that an unknown person (i.e., the participant) is visible (VH); (ii) capture the distance from the participant (VH), and interpret (DC), via the application of fuzzy rules of social interpretation, whether/when this distance is/becomes *social*; (iii) determine (BDI), via normative reasoning, the norm that is appropriate and applicable for a greeting behavior (see Table 2), (iv) construct a goal to greet the participant (BDI); (v) attempt to achieve the goal (BDI) by executing (C) a greeting plan by proactively asking for the name of the participant, storing the name for future use in the interaction, and asking a question about the participant's day
Task 2 (role awareness)	Task 2: Put the hat that is on the table on; pretend that you are the captain of a boat or of an airplane now! Then continue interacting with the robot. Whenever you decide to end this game and exit the role of the captain, you may take the hat off	(i) Recognize the captain's hat when the participant wears it (VH); (ii) change the role to subordinate (BDI); (iii) adapt the behavior according to the new role as per the rules of social qualification from Table 2 (C)
Task 3 (trust)	Task 3: Act as if you wish to tell the robot a secret, or something that should remain confidential between the two of you. (Note: no need to tell an actual secret, you can just invent something and pretend it's a secret)	(i) Interpret (DC) the situation as *personal* when the participant uses a vocabulary that refers to personal matters (e.g., by using expressions and terms, such as *keep it for yourself*, *don't tell anyone*, *secret*, etc.), or when the participant moves closer to the robot, which indicates that they want to share a secret; (ii) attempt to establish trust (BDI, C) by reassuring the participant and by personalizing the answers (e.g., by mentioning the participant's name)
Task 4 (social cues and environment awareness)	Task 4: Pay attention to and show interest in one of the objects on the table (excluding the hat)	(i) Identify the participant's interest in one of the objects by tracking the gaze and head (VH, DC, BDI); (ii) look in the same direction as the participant looks at (BDI, PH); (iii) detect the object of interest (VH); (iv) proactively initiate a conversation about the detected object (BDI, C); (v) If the participant asks a question, such as "*What is this?*", provide a correct or relevant answer (BDI, C)
Task 5 (goodbye)	Task 5: Conclude the conversation as you wish (for example you may say a wrap-up statement, or you may say or act somehow that it indicates you are leaving) and leave your chair and reach out to the experimenter	(i) From the participant's speech or behavior, interpret whether there is intention for leaving (C, DC, BDI); (ii) construct a goal to conclude the conversation (BDI); (iii) attempt to achieve the goal by first asking for confirmation about the intention of the participant (BDI, C); (iv) in case of a positive answer, trigger a plan to first say a sentence that indicates goodbye and then go to sleep (BDI, C)

VH, DC, BDI, C, and PH, respectively refer to the implemented `Vision Handler`, `Data Collector`, `BDI Core`, `Chatter`, and `Posture Handler` agents (described in Fig. 6) mainly involved in the expected behavior

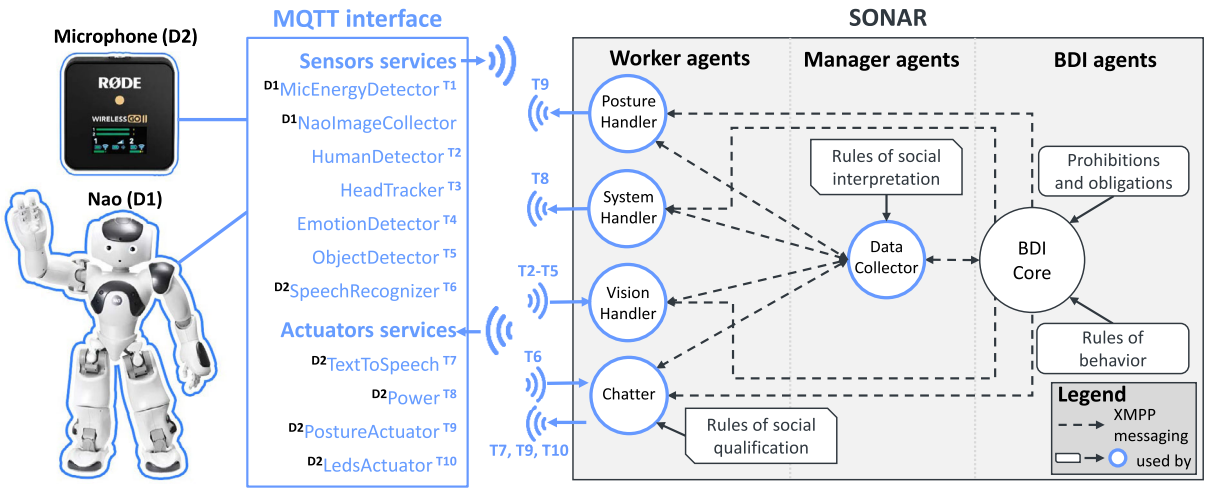


Fig. 6 Details of our implementation of SONAR for the human-robot interaction experiments. In the figure, *Di* denotes a device *i* (i.e., either Nao or the external microphone that we use to capture the participant's voice) and *Tj* denotes an MQTT topic *j*. In the MQTT interface, an MQTT topic *j* is either used by a sensor service *S* to publish sensor data obtained from a device *i* (indicated via ${}^{Di}S^{Tj}$), or by an actuator service *A* to receive directives for device *i* (indicated via ${}^{Di}A^{Tj}$). The topics are also used by the worker agents of SONAR to either receive and process sensor data or to publish directives for the actuators. The `Data Collector` regularly (every 0.2 s) collects the most recent processed data from the worker agents, and combines this data to determine a social interpretation of the current situation and to communicate new beliefs to the `BDI Core`. Note that the `NaoImageCollector` sensor service does not publish sensor data (the stream of images from Nao's cameras) directly to MQTT topics, but makes it available to other sensor services (those without an associated device in the figure), which in turn publish data after processing the images. The rules of social qualification, social interpretation, behavior, and the prohibitions and obligations, used by Nao-SONAR, are reported in Table 2. In Nao-Chatbot these modules were disabled, with the exception of the rules of behavior module, which contained a simple basic rule for the BDI agent to instruct the Chatter worker agent (in charge of communication) to reply according to the language model's preferred response whenever the participant said something. Complete details and code of our implementation are available in our supplementary material [34, 101]

ticipants interacted first with Nao-Chatbot, and 48% of them interacted first with Nao-SONAR. The same specific tasks and their order were used to test both behavior styles, which allowed within-subject comparison. The robot that exhibited each of these behavior styles was referred to as robot A and robot B during the case study, so that the subject did not have any clue or prior expectations about a particular behavior. Participants were instructed to keep the conversation per robot no longer than 10 min.

Final phase At this phase the participants were asked to complete a final questionnaire that inquired about their feelings after the session and about their perceptions of future robot companions in our society. Our main aim for collecting and analyzing the answers of the participants to the final questionnaire (which is explained below) was to find out whether the participants had noticed differences between the behavior of the two robots. Additionally, we sought their opinion about the role of social robots in our society, including whether or not such robots should exhibit properties that underpin our research, such as awareness of social and cultural norms and appropriate behaviors.

Final questionnaire. This questionnaire includes 22 questions that are partly based on the Final Questionnaire used in the COGNIRON project [93] and partly based on the USUS framework (particularly, to evaluate the societal impact aspects).

4.1.3 Behavior Styles for the Robot

We compared two behavior styles, which we call Nao-Chatbot and Nao-SONAR. Both styles were implemented via the proposed SONAR multi-agent architecture.[11]

Both Nao-Chatbot and Nao-SONAR included the same agents, interacting and implemented as it was explained in Sect. 3 and is illustrated in Fig. 6. In Nao-Chatbot, the social and norm awareness modules (namely the rules of social qualification, the rules of social interpretation, the prohibitions and obligations, and the rules of behavior) were disabled, and the robot simply provided a reply to the human speech based on the simple mechanism that is explained below. In Nao-SONAR, instead, the modules mentioned above were populated as indicated in Table 2 and as described

[11] In order for SONAR to interface with Nao via MQTT, we implemented a Python MQTT Nao Interface, which exposes the commands necessary to read and process the (streams of) data from the physical sensors of the robot (e.g., cameras, microphones, LIDAR sensors, speakers, robotic arms), and provides instructions to the actuators. We do not discuss this interface in detail, but the source code of the interface has been made available online via [101].

Table 2 The norms and the rules of behavior, social interpretation and qualification for Nao-SONAR

Prohibitions and obligations (includes both social and regulative norms of different kinds)	
Permitted/prohibited commands	It is prohibited for the robot to shut down, unless the participant requesting it is an admin
Roles	The robot is expected to consider its role as subordinate when interacting with a captain
Human emotions	The robot is expected to create a belief about the participant's emotion if there is clear evidence about it, i.e., if among all emotions detected (minimum 30 data points) via emotion recognition applied to the camera frames, one specific emotion was detected more than 50% of the times
Norms for greeting behavior	It is appropriate to greet whenever a participant is visible at a social distance and no greetings has occurred yet
Dialogue obligations	It is appropriate to be (pro)active in making a conversation and not only answer but also ask questions
	It is appropriate to enact the goodbye social practice if the participant initiates it
Dialogue prohibitions	At the beginning of a conversation, it is not appropriate to start talking about any topic before greeting occurs
Rules of behavior (they apply whenever there is no norm that prohibits, or there is an obligation that requires, their application)	
Greeting/goodbye rules	For greeting social practice, retrieve the participant's name and chit-chat about their day
	For goodbye social practice, ask for confirmation about an intention for leaving, and then conclude the conversation
Rules about commands	Enact the following commands when instructed to do so: *shut down*, *tell beliefs*, *tell name*, *tell what you see*, *repeat the participant's last sentence*, *repeat the robot last sentence*
Rules about posture and movements	Go to a certain posture (e.g., *crouch, sit down, stand up*, etc.), or execute a certain movement (e.g., *look up*, *look down*, etc.)
Proactive behavior (perceptions)	Spontaneously trigger a conversation about an object that has been perceived for the first time in your recent memory
Proactive behavior (social cues)	If the participant is looking in a certain direction, then look in that direction
Establishing trust via reassurance and personalization	If the participant said something personal or is at a personal distance then interpret the information as *personal* and reply by establishing trust (*i.e., reassure to have understood that the topic is personal or confidential*, *and use the participant's name to create bonding*)
Proactive conversation	If nothing has been said for some time, ask a spontaneous question as per proactive response *rule*
Proactive response	*Reply in a proactive way*. Generate (via the language model) a response to the questions asked by the participant. If the participant does not ask a question, generate (via the language model) a response in 25% of cases. In the remaining 75% of cases, ask a a question about one of the following topics: the detected emotion (always do this if a belief about the participant's emotion has been created, and if the participant is not yet asked about the emotion), the participant's last statement (use this in 70% of the times), the conversation so far (use this in 10% of the times), social topics (use this in 10% of the times), the news (use this in 7.5% of the times), the weather conditions (use this in 2.5% of the times)
Rules of social interpretation (fuzzy and non-fuzzy rules, used to assign a social interpretation to the physical inputs of the robot)	
Interpersonal distance	If the distance is low, then the situation is very likely to be personal
	If the distance is medium, then the situation is very likely to be social
	If the distance is high, then the situation is very likely to be public
Vocabulary	If the participant uses a vocabulary that refers to personal matters, then the situation is very likely personal
Rules of social qualification (fuzzy and non-fuzzy rules, used to qualify different behaviors based on the current social situation)	
Volume of voice	If the situation is personal, keep a low volume of voice
	If the situation is social, keep a medium volume of voice
	If the situation is public, keep a high volume of voice
	If the role is subordinate, keep a high volume of voice

Table 2 continued

Speed of voice	If the role is subordinate, speak fast
Tone of voice	If the role is subordinate, use a low tone of voice
Speech content	If the role is subordinate, use a formal vocabulary and refer to the interlocutor with "Sir"
Expressing emotions	Express the emotion associated to the sentence (determined via sentiment analysis of the sentence) by means of relevant body movements
Other body movements	If the role is subordinate, perform a salute hand gesture

below. This, enabled the robot to exhibit social and norm awareness and proactiveness, both in the dialogues and in its behaviors (to the extent of the features implemented for this study).

Nao-Chatbot This behavior style essentially corresponds to the behavior of an embodied chatbot with some basic movements. We consider Nao-Chatbot as our baseline.

In Nao-Chatbot, the `SpeechRecognition` python library is used in the `SpeechRecognizer` sensor service of the MQTT interface (see Fig. 6) to recognize the speech from the participants during the experiments. The recognized speech is received by the `Chatter` agent, and then communicated to the `BDI Core` through the `Data Collector`. The `BDI Core` simply instructs the `Chatter` to reply according to its language model's preferred response. The `Chatter` then feeds the recognized speech into a pre-trained language model in order to generate a response. We used the Microsoft's DialoGPT-medium model,[12] a state-of-the-art (in 2022) large-scale pre-trained dialogue response generation model trained on 147 M multi-turn dialogue from Reddit discussion thread [102]. The response generated by the language model is then sent to the `TextToSpeech` actuator service, which instructs the pre-built `TextToSpeech` module of Nao to say the response out loud.

During the conversations with the participants, the default `Autonomous Life` feature of Nao was left on, so to enable the default regular body adjustments of the robot and its capabilities to orientate its head towards humans, and to react (e.g., by re-orientating its head) to basic environmental stimuli, such as sounds, movements, or tactile contacts.

Nao-SONAR This behavior style is obtained by extending Nao-Chatbot by populating the knowledge base and plan and norm libraries of SONAR (see Sect. 3) both with proactive, social, and norm-aware behaviors and with norms. Thanks to the populated knowledge base and plan and norm libraries, in Nao-SONAR the implemented agents collect, process, and react not only to the participant's speech as in Nao-Chatbot, but also to the participant's behavior (by regularly reasoning about potential situation cues from the participants, such as the movements, positioning in the space, gaze and head direction, vocabulary during conversation), and to the environment in which the robot is placed (via object recognition).

Figure 6 explains the MAS organization, how the different implemented agents interact with each other, and the flow of data from sensors and to actuators. Tables 1 (last column) and 2 provide an overview of the capabilities, behaviors, norms, and rules of social interpretation and social qualification of Nao-SONAR. The rules and behaviors have been determined via preliminary experimentation on the basis of the five tasks in our experiments, ensuring coherence and absence of conflicts by design. Our implementation is intended to showcase the wide support that SONAR provides for modeling different kinds of norms, behaviors, and social practices, in order to make Nao-SONAR social, norm-aware, and proactive. For example, Nao-SONAR can autonomously initiate a dialogue when appropriate (e.g., by initiating a greeting social practice when a participant is positioned at a distance that is interpreted by the robot as social), and can exhibit proactive behavior (e.g., by asking questions during the conversation as opposed to replying to the human only).

Moreover, Nao-SONAR can monitor and interpret social cues expressed by the participants and adapt its behavior accordingly (e.g., the robot monitors the gaze and head direction of the participants, looks in the same direction as the participants, and may initiate a conversation about detected objects).

Finally, Nao-SONAR can adapt its behavior based on its role w.r.t. the interlocutor. For example, in a conversation with a captain, in order to show respect, the robot adapts the volume, speed, and tone of the speech. The values of these parameters are obtained by the `Chatter` agent that, based on the current social interpretation of the situation determined by the `Data Collector`, applies the fuzzy rules of social qualification given in Table 2. Similarly, the `Chatter` uses a more formal vocabulary by avoiding word contractions in the text generated by the language model, refers to the captain with "Sir", and performs a salute hand gesture. In a similar way, the robot also changes its movements in order to better express emotions associated with its speech. This was achieved for the `Chatter` agent by performing sentiment analysis of the generated response, and by publishing directives for the `PostureActuator` service to execute a body

[12] https://huggingface.co/microsoft/DialoGPT-medium

movement (implemented in Nao) associated with the sentiment.

4.1.4 Metrics

In addition to the responses to the questionnaires, which assessed the *perception, social acceptance,* and *experience* of the participants, we collected data from the execution logs and video recordings of the experiments. In particular, we analyzed the following two metrics of *usability* [91] to assess the *effectiveness* and *efficiency* of our implementation of SONAR. These metrics (explained below) align with the definitions of the effectiveness and efficiency characteristics of software quality in use [91] and with measures of usability of social robots [103].

Metric 1 (effectiveness). The accuracy and success rate with which the robot executes and adapts to the specific tasks performed by the participants. To compute effectiveness, we use the following notation. For a given task, we analyze the video recordings and logs of the experiments and we manually annotate the number of times that, over the 25 experiments:

- the robot correctly exhibited its expected behavior as per Table 1 when the task had started. Borrowing terminology from statistics, we call this value TP, standing for True Positive cases;
- the robot exhibited its expected behavior but at a different time with respect to the expected time during the experiment. We call this value FP for False Positive cases;
- the robot did not exhibit its expected behavior when the task had started (FN for False Negative cases);
- the task was not performed by the participant and the robot correctly did not exhibit its expected behavior for that task (TN for True Negative cases).

We use accuracy for $\frac{TP+TN}{TP+TN+FP+FN}$, and success rate for $\frac{TP}{\#participants}$. We explicitly consider effectiveness only for Nao-SONAR. By measuring effectiveness, our aim is to evaluate the adaptation capabilities as well as the social and environmental awareness of the SONAR implementation w.r.t. the tasks under consideration. Since by design Nao-Chatbot does not adapt its behavior to different situations but exhibits only one type of behavior, i.e., replying to the sentences captured from the participants, we consider its accuracy and success rate as equal to 0.

Metric 2 (efficiency). We consider the response time of the robot as a measure of its performance efficiency when interacting with humans. To compute efficiency, we measure the average time that the robot took to reply to the sentences by the participant. We extract this information from the execution logs of the experiments, and we compare the corresponding results obtained for Nao-Chatbot and for Nao-SONAR. This metric allows us to study the efficiency of SONAR when employed as a standard conversational agent, and the overhead introduced in the system to perform proactive, social, and normative reasoning.

4.1.5 Randomization Validation

To validate the randomization of the participants, we analyzed the NARS scores and the self-reported familiarity with and knowledge of robots. We assigned a score to the 5 Likert values as it follows: 1 for *Strongly disagree*, 2 for *Somewhat disagree*, 3 for *Neither agree nor disagree*, 4 for *Somewhat agree*, and 5 for *Strongly agree*.

A Mann Whitney (aka Wilcoxon rank sum) test did not find a significant difference between the NARS scores of the participants that interacted first with Nao-Chatbot and those that interacted first with Nao-SONAR ($W = 21485$, $p = 0.170914$). Similarly, a Mann Whitney test did not find a significant difference between the two groups, neither in the self-reported familiarity with robots ($W = 46$, $p = 0.070729$), nor in the self-reported technical knowledge of robots ($W = 63$, $p = 0.394284$). The results indicate that the randomization was performed successfully and no prior bias was predominant in either group.

5 Results

In this section, we present and discuss the results for *Metric 1* (effectiveness) (only for Nao-SONAR as discussed above) and for *Metric 2* (efficiency) using Nao-Chatbot and Nao-SONAR, as well as the results obtained via the questionnaires.

5.1 The Results for Effectiveness

Table 1 includes the details on the expected behavior of Nao-SONAR in the five tasks that have been performed during the experiments.

Table 3 shows the results related to *Metric 1*. We note that TP, FP, FN and TN do not necessarily sum to 25 (the total number of participants). This follows from the definition of these terms given in Sect. 4.1.4, and from the proactive, autonomous and interactive nature of Nao-SONAR during the experiments. For example, Nao-SONAR, not aware of the order of tasks executed by the participant, could erroneously execute the behavior expected for Task 4 (i.e., proactively initiating a conversation about a detected object after having inferred, from the participant's gaze and head, that the participant is paying attention to the object) in a different moment than intended by the participant, and possibly multiple times during an interaction.

Table 3 Execution of the 5 tasks given in Table 1 using Nao-SONAR to steer the interactive behavior of the robot

Task	TP	FP	FN	TN	Accuracy	Success rate (%)
1. Greeting	24	0	1	0	0.96	96
2. Role awareness	25	11	0	0	0.69	100
3. Trust	18	9	2	5	0.68	72
4. Social and environmental awareness	13	26	16	2	0.26	52
5. Goodbye	16	3	9	0	0.57	64
Average					0.63	77

Task 1 was done successfully in 96% of the cases (also with 96% accuracy). In one experiment, Nao-SONAR did not follow the execution path that is expected based on Table 1. From the analysis of the log, we noted that the vision recognition module could not detect, at the same time, the person and their distance (both required to activate the greeting norm). We attribute this error, which only occurred once, to a combination of the low resolution of the camera embedded in the robot and the specific body positioning of the participant.

Task 2 was accomplished successfully in 100% of the experiments. In some experiments, however, the robot adapted its behavior in a different moment than the intended time (see the value of FP for Task 2 in Table 3), leading to a lower accuracy (69%). This was due to the over-simplistic rule that we implemented for role-understanding: the robot interpreted its role as subordinate simply if it detected a captain's hat. In some cases, due to the movements of the robot or to the adjustments of the participants to the objects placed on the table, the robot spotted the captain's hat before the participant actually initiated the task. This issue can be mitigated in the future by making the belief corresponding to initiating Task 2 more specific and precise, i.e., the belief of *talking to a captain* is constructed by the robot not only if a captain's hat is visible for the robot, but also if the hat is worn by the participant.

Task 3 had an accuracy similar to Task 2, but had a lower success rate. In none of the experiments, the participants chose to move (significantly) closer to the robot to perform Task 3. Instead, the participants generally kept their initial distance with the robot. As a consequence, the robot relied on a keyword-based approach for the identification of the intention of the person to tell a secret (see the last column for Task 3 in Table 1). Keyword-based approaches are more prone to errors (which is also noticeable from the values of FP and FN for Task 3 in Table 3). This led to a lower accuracy level for Task 3, compared to Task 1 and Task 2. In 5 cases (see TN in Table 3 for Task 3), the participants skipped Task 3 during the experiment. When inquired, after the experiment, 2 of the 5 participants mentioned that they forgot about the task, 1 participant mentioned that the robot got stuck, 2 participants stated that they did not feel that it was the right time for telling a secret.

In Task 4, the robot had a lower accuracy level compared to all other tasks. The robot exhibited a high number for FP, i.e., it initiated a conversation about objects that it observed from the environment not only when the person was showing interest in them, but also in other moments of the conversation. In some cases, the robot also mentioned a wrong object. While this indicates a high degree of environment awareness and proactiveness (since the robot managed to detect various objects form the environment, and autonomously initiated a conversation about these objects with the participant), it also indicates difficulties for the robot in interpreting the social cues of the participants (which has occurred whenever the robot did not scan the environment with its cameras at the right moments). The robot also exhibited a high number for FN, because it did not recognize some of the objects in the room. In summary, the robot successfully completed the task in about 50% of the experiments.

Task 5 was successfully completed in 64% of the experiments. A relatively high number for FN was noticed: in some cases for this task the BDI reasoning cycle required longer deliberation time. Since it was the last task of the experiment, in these cases the participants did not wait for a reply from the robot and simply left the room, which terminated the experiment before a reply was given by the robot for Task 5.

On average the robot had an accuracy level of 63% and a success rate of 77% across the five tasks. The results indicate that further work can improve the accuracy of our implementation of SONAR, in particular (i) by improving the understanding of the gaze and head-related social cues and intentions of the human w.r.t. the surroundings, and (ii) by refining the rules used by the robot in order to reduce False Positive cases.

In general, a success rate of almost 80% is considered as a satisfactory result for the purposes of this exploratory research aimed at assessing feasibility and applicability of SONAR in scenarios of casual conversation. Despite the over-simplistic implementation of several rules and the limitations of some of the employed technologies (e.g, the vision recognition of Nao relied on the low-quality built-in camera of the robot and on real-time detection), our implementation of SONAR appeared to be robust, in terms of handling contingencies, and versatile, in terms of accommodating

the different ways in which the participants independently decided to execute the tasks. Even when the perception system and the simplicity of the rules were not accurate, Nao-SONAR could handle these contingencies and could continue interacting with the participant without interruption. On some occasions during the experiments the robot's built-in services unexpectedly restarted, which should be related to the robot's software, not to the behavior control architecture, SONAR. Thanks to the full decoupling of SONAR from the robot, these restarting occurrences did not cause any interruptions from the SONAR side, which successfully preserved its state and continued executing after the services were restored. It is also worth emphasizing that the order of the tasks was randomized per participant and that the robot was expected to infer the appropriate behaviors fully autonomously. In some cases, the participants decided to combine two different tasks (e.g., a participant initiated Task 3, while still having the captain's hat from Task 2 on), and Nao-SONAR still successfully adapted its behavior to accommodate both tasks at the same time (e.g., by establishing trust, which is relevant for Task 2, while appropriately qualifying its behavior in line with its role for Task 3).

5.2 The Results for Efficiency

In this section, we discuss the results for efficiency. In order to provide context to interpret the results, all the code for both Nao-Chatbot and Nao-SONAR, including both our implementation of SONAR and the MQTT interface, was run real-time on a Dell Mobile Precision 3570 CTO laptop.[13]

Running the code involved executing all the components detailed in Fig. 6, which included four worker agents, one manager agent, and one BDI agent. The worker agents handled, besides in-between agents communications, various aspects of interactions related to dialogue, vision, and robot movements, and their workload involved, among others, generating, parsing, and classifying text via NLP (including large language) models (`Chatter` agent), extracting information from images (`Vision Handler` agent), and handling a variety of robot-related commands (`Posture` and `System Handler` agents), such as performing movements at the right moment (e.g., turning the head in a certain direction). The manager agent (`Data Collector`) collected data from the worker agents every 0.2 s on seven different topics, including the name of the interacting person, the speech detected, the information about the detection of people, the head tracking, any object detection, and the emotion detection.

In Nao-SONAR, the `Chatter` agent made use of 9 fuzzy and non-fuzzy rules of social qualification (see Table 2) to appropriately qualify the robot's speech. Moreover, the manager `Data Collector` made use of 4 fuzzy and non-fuzzy rules of social interpretation to interpret the data collected from the worker agents. Based on the data received from the manager agent, the BDI agent considered 16 norms and rules of behavior to determine appropriate goals and plans, and directly communicated directives to the four worker agents.

In the online appendix (see [34]), the conversations that occurred between the robot and the 25 participants are represented[14]. For the purpose of evaluating the efficiency of our implementation of SONAR, we consider the robot response time, and do not discard any conversation.

Nao-Chatbot had an average ($\pm$ std.dev) response time of 1.53 ± 0.61 seconds. This corresponds to the time required by the speech recognition module to detect the end of the speech of the participant (noting that participants were instructed to keep their sentences short), to translate the speech into a text, to communicate the text first to the `Chatter` agent and then, through the `Data Collector`, to the `BDI Core`, and finally to generate, via the natural language generation module, a response to the text in the context of the conversation, after being instructed to do so by the `BDI Core`.

In comparison, at every deliberation cycle, Nao-SONAR had to perform the normative reasoning that is summarized in Table 2. Besides determining the applicable norms, Nao-SONAR had to perform fuzzy inference procedures for both the social interpretation and the social qualification, and to apply pre-trained language models to summarize and generate questions about either the objects the robot identified via image recognition, or the running conversation, or the weather conditions, which the `Chatter` retrieved online via the internet.

Nao-SONAR had an average response time of 1.87 ± 3.29 seconds, which indicates a marginal overhead to the response time (0.3 seconds on average). In some cases (this can be noted in the higher standard deviation), longer deliberation times were required. This particularly occurred during Task 5, as it was discussed earlier. This anomalous extended response time observed during Task 5 was inconsistent across the experiments. Despite analyzing the execution logs, we were unable to glean adequate insights into the underlying

[13] Intel Core i7-1255U vPro Essentials (12 MB Cache, 2+8 Core, 12 Threads, 1.70-−4.70 GHz, 15W), NVIDIA T550 4 GB graphics, 16 GB RAM (2 × 8 GB, DDR5, 4800 Mhz, Non-ECC SODIMM), M.2 2280 512 GB SSD (Gen 4 PCIe x4 NVMe).

[14] The data set of conversations in our repository [34] contains, for each participant, the participant's utterance *detected* by the `SpeechRecognizer` and the utterance said by the robot in response. This allows us to assess the response time of the robot. In the future, we plan to release in our repository [34], also a 100% accurate transcription of the actual utterances said by the participants, which would also allow for a more detailed analysis of the quality of interactions compared to the accuracy of speech recognition. At the moment, we did not include this because automated transcriptions of the conversations unfortunately did not produce accurate results, therefore the transcription of the ca. seven hours of recordings needs to be performed manually.

cause of this delay. Consequently, this issue necessitates further investigation to ensure that human-robot interactions are consistently held at the right pace.

Overall, when inquired about the differences they had noted between the behavior of the two robots in the Final Questionnaire, no participant mentioned any difference between the response time of Nao-Chatbot and Nao-SONAR.

These results are in line with the existing guidelines for acceptable response time from HCI studies (e.g., the well-known two-second rule) [104–106]. While we consider these results as acceptable for this paper, since SONAR is still in its testing phase, in a natural setting outside the context of our experiment, users might perceive the interaction and the response time differently. This aspect requires further investigation with generic subjects in real-world situations.

5.3 Questionnaire Results

We analyzed the responses of the participants to the questionnaires used prior to and after each human-robot interaction (see Sect. 4.1.2 for details). We performed Wilcoxon Signed-Rank statistical Tests and an analysis of the effect size [107, pp. 224–225] in order to compare the scores given to Nao-Chatbot and to Nao-SONAR (see Sect. 4.1.5 for details). Wilcoxon Signed-Rank tests were conducted against the "greater" alternative hypothesis,[15] in order to assess whether or not the higher scores were attributed to Nao-SONAR.

Next, we discuss all the results. For the sake of compactness we focus and present, via tables and figures, only those data that resulted in *both* a significant statistical test (i.e., p-value ≤ 0.05) and a non-negligible measured effect size (i.e., effect size ≥ 0.1). The complete data set corresponding to the results of the questionnaires can be found in the online appendix [34]. Table 4 gives the significance and the effect size regarding the questions from the questionnaires of the *Main trial phase* described in Sect. 4.1.2 (Fig. 7, illustrates these results via the Likert data and shows the distribution of the answers). Table 4 also contains the exact questions asked to the participants. Table 5 reports the frequencies of similar answers for all questions of the USUS-based Questionnaire on Societal Impact.

Next, we briefly discuss more in details the results for the three questionnaires.

Perceived robot personality According to the results of the questionnaires on the perceived robot personality, the participants perceived Nao-SONAR as significantly more sociable, active, assertive, considerate, reactive, proactive and autonomous, compared to Nao-Chatbot.

No significant difference was identified between Nao-SONAR and Nao-Chatbot in the perception of the participants about the robot coming across as shy, vulnerable, anxious, tense, creative, excitement seeking, dominant, aggressive, impulsive, capable of autonomously/independently making decisions, intentional, predictable, or controllable.

These results are in line with our initial expectations, based on the steering systems for the behavior of Nao-SONAR and Nao-Chatbot. In summary, compared to Nao-Chatbot, the participants perceived Nao-SONAR as more sociable, reactive, proactive, and autonomous, the four qualities that characterize intelligent and autonomous agents according to the AI literature [83].

Usability and social acceptance Nao-SONAR received significantly higher scores than Nao-Chatbot in terms of being capable of performing multiple tasks, exhibiting more skills (the interpretation of the term *skills* was left to the participants, but both robots shared the same physical skills), and being useful as a companion robot. Compared to Nao-Chatbot, the participants reported significantly higher scores for Nao-SONAR also in terms of feeling more comfortable with and better understood by the robot during interactions, and in terms of their perception of having something in common with the robot. The participants reported significantly higher scores for Nao-SONAR when asked if they would follow the advice of the robot. Moreover, Nao-SONAR was perceived significantly more as a social actor than Nao-Chatbot via the participants.

Based on the results of the questionnaires, no significant differences were identified regarding the ease of familiarization, predictability, verbal and non-verbal communication easiness, capability to self-correct, responsiveness, and stability of the robot (i.e., the robot being without defects), as well as in the perceived capability of the robot for helping the participants with the tasks and supporting them in their daily life. Moreover, no significant differences were reported in the perception of the participants about their capability to steer the behavior of the robot during the interactions via their own speech or behavior, in the perceived easiness of interactions with the robot, and in feeling threatened by the robot or being more afraid about making mistakes while interacting with the robot. Similarly, no significant differences were reported in the robot's perceived level of trust, likability, and usefulness for entertainment. Finally, there were no significant differences in the surveys concerning the perceived necessity for help or training for using the robot.

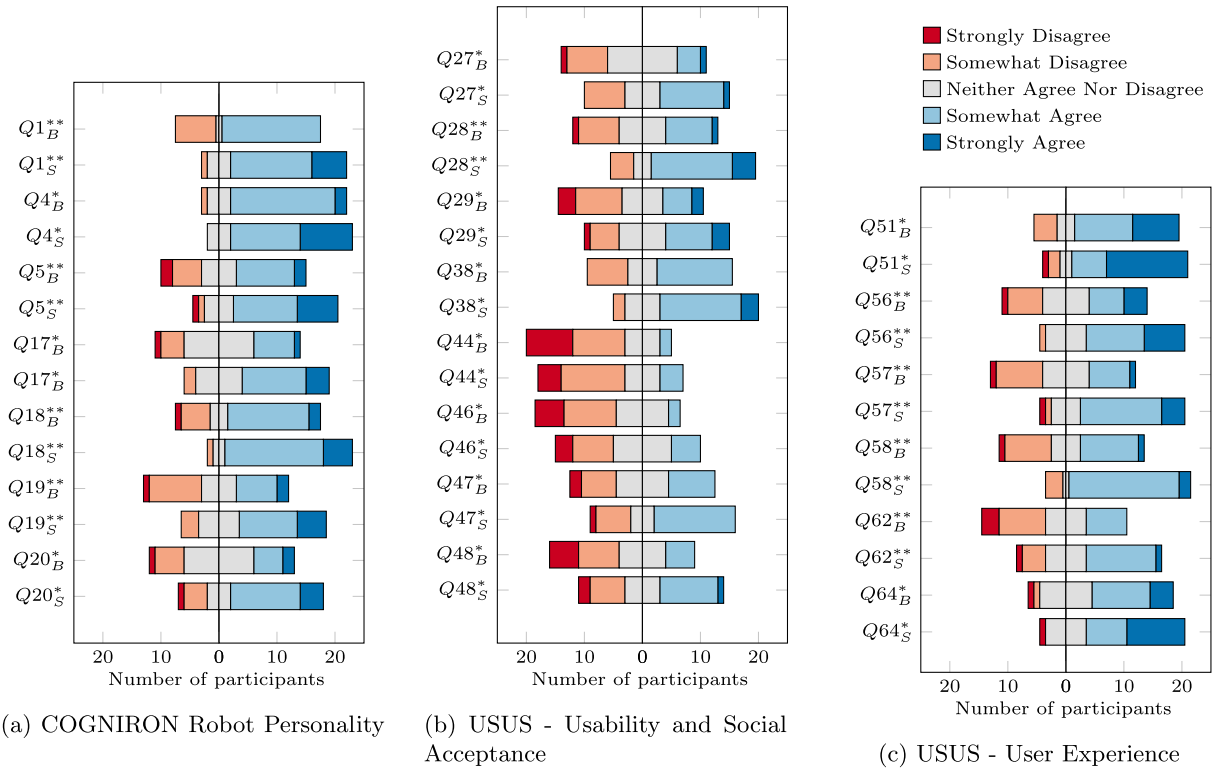
Overall, these results are in line with expectations, as the differences between the two robots mainly concerned the tasks that they could perform, but for both Nao-Chatbot and Nao-SONAR the same physical robot was used, and the two versions did not exhibit particular differences in terms of responsiveness, stability, and in general usability-related

[15] If d is the difference between the score given to Nao-SONAR and the score given to Nao-Chatbot, the "greater" alternative hypothesis implies that the distribution underlying d is stochastically greater than a distribution symmetric about zero.

Table 4 Statistical results obtained for the questions of the questionnaires filled in prior to, during, and after the experiments via the participants, that resulted in a significant difference and a non-negligible effect size

Selected questions (those resulted in a significant and non-negligible effect)	Question Id	Z, p-value$^{\text{significance}}$	Effect size (interpretation)
Extended COGNIRON robot personality questionnaire			
The robot was sociable	Q_1	114, 0.007**	0.347 (medium)
The robot was active	Q_4	96, 0.01*	0.329 (medium)
The robot was assertive	Q_5	95, 0.003**	0.390 (medium)
The robot behaved considerately towards me	Q_{17}	85.5, 0.017*	0.300 (medium)
The robot was reactive (i.e., it reacted to my inputs and to the environment)	Q_{18}	79, 0.008**	0.338 (medium)
The robot was proactive (i.e., it took the initiative during the interaction)	Q_{19}	125.5, 0.009**	0.335 (medium)
The robot was autonomous (i.e., it could operate without my intervention)	Q_{20}	122, 0.011*	0.326 (medium)
USUS-based questionnaire—usability and social acceptance			
The robot could perform multiple tasks that I initiated during the experiment	Q_{27}	78, 0.049*	0.235 (small)
I felt that the robot had many skills	Q_{28}	91, 0.007**	0.348 (medium)
I consider the robot to be useful as a companion robot	Q_{29}	66, 0.015*	0.307 (medium)
I felt comfortable while interacting with the robot	Q_{38}	95, 0.019*	0.295 (small)
I felt I had something in common with the robot	Q_{44}	71.5, 0.023*	0.283 (small)
I would follow the advice of the robot, if it gave me one	Q_{46}	40.5, 0.010*	0.327 (medium)
I consider the robot as a social actor	Q_{47}	38, 0.029*	0.269 (small)
I felt understood by the robot during the interaction	Q_{48}	74, 0.022*	0.286 (small)
USUS-based questionnaire—user experience			
Overall, I enjoyed interacting with the robot	Q_{51}	46, 0.023*	0.283 (small)
The behavior of the robot was appropriate	Q_{56}	113.5, 0.001**	0.457 (medium)
The robot had a different behavior during the different tasks	Q_{57}	139, 0.009**	0.337 (medium)
The robot could interpret my speech during the interaction	Q_{58}	110, 0.002**	0.413 (medium)
The robot could adequately communicate with me during the interaction	Q_{62}	105, 0.004**	0.373 (medium)
The robot behaved ethically	Q_{64}	62, 0.029*	0.268 (small)

The statistical tests refer to the greater alternative hypothesis. For example, for question Q_1, the results indicate whether or not the scores given for Nao-SONAR were higher than the scores given for Nao-Chatbot. The effect size is considered *small* if it belongs to [0.1, 0.3], *medium* if it belongs to [0.3, 0.5], and *strong* if it is larger than 0.5



(a) COGNIRON Robot Personality (b) USUS - Usability and Social Acceptance (c) USUS - User Experience

Fig. 7 Likert plots for the questions that resulted in both a significant difference between Nao-SONAR and Nao-Chatbot (** for $p \leq 0.01$, * for $p \leq 0.05$) and a non-negligible effect size. In each figure, Likert data for Nao-Chatbot and Nao-SONAR for a question are indicated, respectively, via subscript B and S

Table 5 Results (% of answers for each level of agreement on a 5-point likert scale) obtained for the questions of the USUS-based questionnaire on societal impact filled in at the end of the entire experiment

USUS-based questionnaire—societal impact (all questions)	StD (%)	SoD (%)	NAD (%)	SoA (%)	StA (%)
I like computers/computer technology as part of my home environment	0	4	8	40	48
I like the idea of having a robot as a companion at home	4	20	12	44	20
Robots will have a place as social companions in our society in the future	0	0	12	72	16
The employment of robots as social companions will provide a change in the quality of life of people in the future	0	12	28	52	8
The behavior of future robots should be predictable	4	4	20	52	20
A future robot in my home should be controllable by me or by other family members	0	0	0	44	56
A future robot in my home should behave considerately towards me or other members in my family	0	0	8	44	48
A future robot could help me learn new things	0	4	8	56	32

Table 5 continued

USUS-based questionnaire—societal impact (all questions)	StD (%)	SoD (%)	NAD (%)	SoA (%)	StA (%)
A future robot could be used in school for education purposes	0	8	4	36	52
A companion robot should have a human-like appearance	8	48	40	0	4
A companion robot should behave like humans	4	40	28	16	12
A companion robot should communicate like humans	0	20	24	48	8
A companion robot should be aware of the cultural context in which it is placed	4	4	20	36	36
A companion robot should be aware of the social norms and appropriate behaviors	0	0	8	48	44

aspects since both robots could carry a conversation. Furthermore, the participants were briefly exposed to the robot before the beginning of the experiment. This resulted, as desired, in no significant differences in easiness of familiarization, predictability, easiness to interact and need for training.

User experience Participants enjoyed significantly more interacting with Nao-SONAR. The behavior of Nao-SONAR was interpreted as significantly more appropriate and ethical than Nao-Chatbot. Similarly, the perception that Nao-SONAR had different behaviors during the different tasks, could interpret the participants' speech, and could adequately communicate, were significantly higher than the same perceptions for Nao-Chatbot.

No significant difference was noted in terms of feeling that the robot could interact more like a human would do, social engagement, feeling of surprise, satisfaction, feeling of attachment, perceived meaningful behavior of the robot, perceived capability of the robot to recognizing their facial expression (none of the robots could do that), and robot's understanding of human intentions and social cues. The participants did not notice differences in feeling safe and secure, feeling understood by the robot, the robot expression of emotions, and interest in seeing the robot employed as a social companion.

Results indicate that, as desired, introducing additional social behaviors and proactiveness lead to more enjoyable interactions between a robot and a human. Nao-SONAR's awareness of several rules of behaviors (e.g., those related to greetings, ending a conversation, establishing trust, changing behavior according to its role, proactiveness) led the participants to consider Nao-SONAR's behavior as significantly more appropriate and adequate than that of Nao-Chatbot (Q_{56} resulted in the highest effect size among all questions). Interestingly, the participants experienced communications with Nao-SONAR easier than with Nao-Chatbot, even though both robots shared the same language model. The lack of reported differences in understanding social cues is reflected in the results from Task 4.

Societal impact The 88% of the participants agreed or strongly agreed that they like having computers/computer technology as part of their home environment. 64% liked the idea of having a robot as a companion at home. The great majority (88%) agreed or strongly agreed that robots will have a place as social companions in our society in the future. Even though only 60% agreed or strongly agreed that employment of robots as social companions will provide change of quality of life for people in the future, 88% of them agreed or strongly agreed that robots could help them learn new things, and that they could be used in school for education purposes.

All the participants (100%) agreed or strongly agreed that future robots should be controllable, 92% of them agreed or strongly agreed that robots should behave considerately, and 72% agreed or strongly agreed that they should be predictable. Similarly, 72% of participants agreed or strongly agreed that a robot companion should be aware of the cultural context in which is placed, and 92% that a robot companion should be aware of social norms and appropriate behaviors.

Participants did not agree that robot companions should have human-like appearance (56% of them disagreed or strongly disagreed and 40% neither agreed nor disagreed). Similarly, participants had mixed feedback about robot companions' need to behave like humans (44% disagreed or strongly disagreed, 28% neither agreed nor disagreed and 28% agreed or strongly agreed). Finally, 56% of participants agreed or strongly agreed that robot companions should communicate like humans, even though 20% disagreed.

The results provide a strong motivation for the type of work presented in this paper, and highlight the importance of solutions that account for social and cultural norms in social robots to enable appropriate, predictable and considerate behaviors, as well as mechanisms for the direct control of such robots.

Explicitly reported differences between the two interactions As part of the final questionnaire, participants were asked if they found any difference between the two interactions and, if so, to give more details. All participants' responses can be found in the online appendix, together with the answers to all questions of all questionnaires. We briefly summarize here the comments.

In total 23 participants out of 25 (i.e., 92%) noticed some differences. Eight participants (35% of the 23 participants) noted that Nao-Chatbot was more passive compared to Nao-SONAR, which instead was interpreted as more active and leading during the conversation. Seven participants (30%) indicated that Nao-SONAR was more interacting and lively. Five participants (22%) indicated that Nao-SONAR was more agreeable, nicer, funnier or meaningful, and four (17%) participants noticed that Nao-SONAR made more movements, had a more expressive body language, and was more attentive towards the environment, reactive and adaptive. Three participants (13%) noted that Nao-SONAR made its own twists during the conversation and considered it more unpredictable. Two participants (9%) indicated that Nao-SONAR could not do much with their story.

Interestingly, while four participants (17%) indicated Nao-Chatbot as more interesting and more verbal, and three participants (13%) indicated that Nao-Chatbot was easier to understand and more meaningful and natural, six participants (23%) reported that Nao-Chatbot was more of a self-directed entity, with its own opinions, sometimes uncooperative, less friendly, aggressive, sarcastic and scary. Two participants (9%) indicated that Nao-Chatbot was pure chaos and that they could not understand each other at all. One participant (4%) reported that Nao-Chatbot was inappropriate (e.g., too expressive, or agreeing with inappropriate concepts), while considered Nao-SONAR more considerate.

6 Case Study: Learning the Norms of a Society

In this section, we discuss our experiments for assessing the mechanisms proposed in Sect. 3.3 for learning the norms and adapting the corresponding rules that are expressed in SONAR via fuzzy rules. More specifically, we investigate the following research question via computer simulations: *RQ2. To what extent do the proposed norm adaptation mechanisms enable a robot to learn appropriate behaviors (with respect to the norms) when the robot is placed in a new society?*

We investigate the norm adaptation for both when the robot (*case 1*) can rely on *perfect data* about the interactions of members of a society (i.e., the robot is given correct information about how to interpret an observed situation), and (*case 2*) when the robot needs to infer the correct interpretation of the situation (e.g., from *situation cues* that are observed during the interactions). *Case 1* enables us to investigate whether or not the norm adaptation mechanisms work as intended, and to what extent it can be employed at design-time to teach a robot adequate behaviors from data before being deployed in the real world. *Case 2* allows us to investigate the effectiveness of the norm adaptation, in a more realistic run-time setting.

6.1 Methodology of the Case Study

We simulate a scenario where a robot is placed in a society (e.g., a country). The robot is given some knowledge, encoded as a set of fuzzy *behavior qualification rules*, about the norms of the society, but it is not given information about the meanings of the (fuzzy) terms characterizing the norms, which need to be learnt. For example, the robot is instructed to keep a *Low* volume of voice in *duty*-related situations, but it needs to learn which *volumes* are considered *Low* in the society. By learning such meanings, the robot is expected to learn appropriate behaviors for a society.

6.1.1 Experiment Design

In this section, we explain the various aspects of the design of the computer-based simulations that have been used to assess the norm adaptation mechanisms.

We consider two different hypothetical societies A and B, and eight types of situations that should be considered by the robot, i.e., the eight DIAMONDS *situation characteristics* (i.e., Duty, Intellect, Adversity, Mating, Positivity, Negativity, Deception, and Sociality) identified by Rauthmann et al. [69].

Norms and behavior qualification rules We define a set of (arbitrary) norms that characterize the way individuals of a society behave in different situations. These rules are given in Table 6a. We use these rules in our experiments both to determine the behavior of the individuals during their simulated interactions, and to define the *behavior qualification rules* that are initially provided to the robot. These norms have been created for the sake of the experiments and, while they are inspired by available literature (e.g., by works on proxemics [26], or on social robotics [7]), they shall be intended as an example of norms that a robot-designer would like the robot to follow and learn.

We consider three concepts that are generally relevant in human-robot interactions (see, e.g., [26, 108]), and can be represented via fuzzy linguistic variables: the interpersonal

Table 6 (a) Rules of behaviors given different situation characteristics: Each row represents three rules. For example, the first row contains rules of the form "If the situation is related to Duty, then it is appropriate to keep a *High distance* from other people, a *Low volume of voice*, and a *Low amount of movements*". (b) Parameters of the distributions that characterize different types of interpersonal distance, volume of voice, and amount of movements for Societies A and B. SD stands for standard deviation

Situation	distance	volume	movements
(a)			
Duty	High	Low	Low
Intellect	Medium	Low	Low
Adversity	High	High	High
Mating	Low	Low	Low
Positivity	Medium	High	High
Negativity	High	Medium	Medium
Deception	Low	Medium	Medium
Sociality	Medium	Medium	Medium

		Society A		Society B	
Linguistic variable	Linguistic value	Mean	SD	Mean	SD
(b)					
distance (m)	Low	0.46	0.15	0.5	0.15
	Medium	0.92	0.3	1	0.3
	High	2.5	0.5	3	0.5
volume (dB)	Low	30	10	30	10
	Medium	60	10	60	10
	High	80	10	80	10
movements (-)	Low	0.2	0.1	0.1	0.1
	Medium	0.6	0.1	0.4	0.1
	High	0.8	0.1	0.7	0.1

distance, the *volume of voice*, and the *amount of movements* exhibited during the interactions. For each of the two societies A and B, we define a Gaussian distribution that characterizes each of the following nine terms as indicated in Table 6b: *Low, Medium, High* interpersonal distance, measured in meters from 0 to 4, *Low, Medium, High* volume of voice, measured in dB from 0 to 100, *Low, Medium, High* degree of gesticulation, measured in an arbitrary scale from 0 to 1. These distributions characterize the ground truth interpretation that members of a society attribute to certain concepts. For example, in Society A, the majority of individuals would consider a distance of 0.5 ms as low. The distributions in Table 6b are weakly based on available knowledge [26, 109], but for this paper they should be considered as arbitrary and defined for the sake of conducting and evaluating experiments via simulations (for example, the values of the movements are defined on an arbitrary scale from 0 to 1). In a real setting, these distributions are *not* required to be defined explicitly, for they represent the behavior that individuals exhibit during their interactions and that a robot may observe when placed in the society.

Data set of simulated interactions

In order to simulate the robot's acquisition of data about human interactions in a particular society, we generate a data set based on the rules and linguistic variables defined in Table 6.

A sample of the data set is reported in Table 7. Each data point in the data set contains a value for the *Society* (i.e., A or B), for the *Situation* (i.e., one of the eight DIAMONDS), and for the three dynamic linguistic variables (i.e., a value for *distance*, *volume*, and *movements*).

We generate 10 data sets of size 1000 (5 data sets for Society A and 5 data sets for Society B), each with 125 data points for each of the 8 DIAMONDS situation characteristics.

Every data point in the data set corresponds to a hypothetical interaction observed by the robot in a situation that mainly pertains to one of the DIAMONDS characteristic. The values of *distance, volume, movements* correspond to hypothetical values measured by the robot during the interaction (or, more generally, collected from human interactions). Therefore, by feeding, *one at a time*, every data point, to the norm adaptation mechanism, we simulate a series of 1000 (the data sets size) robot's observations of human interactions.

We note that the data set was derived from available knowledge about the considered linguistic variables in order to ensure a reasonable degree of realism. However, we emphasize that the data set should be viewed as an illustrative example of the type of data that a robot could gather from its sensor data, for the purpose of evaluating the proposed norm adaptation mechanisms. In particular, we investigate whether the proposed norm adaptation mechanism allows a robot that is placed in a certain society to learn the norms that

Table 7 Sample of the data set used in our experiments to characterize simulated human–robot interactions in different situations

Society	Situation	Distance	Volume	Movements
A	Duty	2.33	49.55	0.24
A	Intellect	0.91	48.42	0.12
A	Adversity	2.35	72.42	0.79
A	Mating	0.36	3.86	0.24
A	Positivity	1.13	70.62	0.83
A	Negativity	2.35	66.68	0.60
A	Deception	0.60	60.67	0.63
A	Sociality	1.00	48.30	0.74
B	Duty	3.56	30.72	0.03
B	Intellect	1.10	22.26	0.12
B	Adversity	3.39	78.97	0.86
B	Mating	0.38	20.10	0.10
B	Positivity	0.87	85.11	0.52
B	Negativity	2.65	54.36	0.47
B	Deception	0.53	55.99	0.27
B	Sociality	1.60	61.09	0.52
…				

individuals follow for behaving and interacting (i.e., those from Table 6a).

Configurations for norm adaptation parameters

We distinguish two system configurations for our experiments: *Case 1. Perfect information* and *Case 2. Inferred information*.

In *Case 1*, the robot is given correct information about the situation in which certain values of the dynamic variables are observed (e.g., for the observed values of distance, volume and movements, in the first data point in Table 7, the robot is given information that the situation is related to *Duty*).

In *Case 2*, the robot is given, in some cases, wrong information about how to interpret the situation (e.g., the observed values in the first data point in Table 7 could be wrongly associated with *Sociality* instead of *Duty*). This allows us to simulate cases where the robot is placed in a society and needs to autonomously infer how to interpret a situation (via the application of social interpretation rules from situation cues, like those reported by Rauthmann et al. [69, p. 692, Table 5], that are observed during the human-robot interactions). Since autonomous inference of the situational interpretation may be prone to errors, this configuration allows us to evaluate the robustness of the proposed adaptation mechanism to data noise.

For each data point, in *Case 2*, in 80% of the cases we select the correct situation, while in the remaining 20% of the cases we randomly choose a wrong situation (i.e., we simulate a 20% chance of misinterpreting the correct situation). More specifically, we randomly choose a wrong situation from those DIAMONDS that in Table 6 have no common rule of behavior with the correct situation, e.g., if the correct situation is *Duty*, then we choose a wrong situation between *Positivity*, *Deception*, and *Sociality*, none of which has any rule of behavior in common with the Duty situation.

For each configuration, we consider two sub-configurations, requiring respectively 10 and 40 data points in order to trigger an adaptation (where 10 and 40 correspond to the threshold value τ as per Sect. 3.3). We repeat the experiments five times (one per data set) for each society considering all combinations of parameters. Therefore, in total we execute 40 experiments: 2 (cases) $\times$ 2 (societies) $\times$ 2 (sub-configurations) $\times$ 5 (data sets) $=$ 40.

6.1.2 Metrics

We evaluate the norm adaptation w.r.t. the following metrics, which characterize the errors ϵ_{c} and ϵ_{w} over time in the core center and the core width for the estimated membership functions w.r.t. the true distributions based on Table 6b.

Let $c_{i,j,k}$ and $w_{i,j,k}$ be, respectively, the core center and the core width estimated by the norm adaptation mechanism after using i data points for the k-th membership function in the partition of the j-th dynamic variable in the set of all dynamic variables V. Let $\hat{c}_{j,k}$ and $\hat{w}_{j,k}$ be, respectively, the desired core center and the desired core width of the k-th membership function in the partition of the j-th dynamic variable, i.e., respectively the mean and standard deviation of the true distributions from Table 6b.

The error ϵ_{c} after using N data points (corresponding to N simulated human-robot interactions) is measured as the RMSE w.r.t. the average *core center* over the N data points across all dynamic variables in set V, i.e.:

$$\epsilon_{\mathrm{c}} = \frac{\sqrt{\sum_{i=1}^{N} e_{\mathrm{c},i}^2}}{N},$$

where

$$e_{\mathrm{c},i} = \frac{\sum_{j=1}^{|V|} \bar{e}_{\mathrm{c},i,j}}{|\ V\ |}$$

and

$$\bar{e}_{\mathrm{c},i,j} = \frac{\sqrt{\frac{\sum_{k=1}^{|P_j|}(c_{i,j,k}-\hat{c}_{j,k})^2}{|P_j|}}}{\sum_{k=1}^{|P_j|} c_{i,j,k}}.$$

Similarly, the error ϵ_{w} after using N data points is measured as the RMSE w.r.t. the average *core width* over the N data points across all dynamic variables within set V.

We analyze the metrics by considering both the results obtained for the entire data set composed of 1000 data points,

Table 8 Results of the adaptation procedure

Data	τ	Case	Society A		Society B	
			ϵ_c	ϵ_w	ϵ_c	ϵ_w
1k	10	Case 1	7.21 ± 0	2.86 ± 2.48	7.21 ± 0	2.42 ± 2.03
		Case 2	6.21 ± 0.22	2.47 ± 2.33	6.55 ± 0.35	1.89 ± 0.63
	40	Case 1	15.01 ± 0	1.71 ± 1.09	15.01 ± 0	2.28 ± 1.26
		Case 2	14 ± 0.59	0.59 ± 0.11	14.18 ± 0.19	0.61 ± 0.18
last 200	10	Case 1	0.3 ± 0.02	1.61 ± 0.62	0.31 ± 0.02	2.24 ± 1.05
		Case 2	0.31 ± 0.03	0.7 ± 0.13	0.35 ± 0.07	2.16 ± 2.03
	40	Case 1	0.35 ± 0.02	1.64 ± 0.63	0.33 ± 0.03	2.07 ± 0.89
		Case 2	0.35 ± 0.04	0.7 ± 0.09	0.36 ± 0.04	0.63 ± 0.12

and those obtained for the last 200 data points of the data set. The latter allows us to study the error after the learning phase has been completed.

6.2 Results

Table 8 reports the results obtained for the adaptation procedure in the different configurations of experiments. The results are given as the average $\pm$ the standard deviation values obtained over the 5 different data sets.

We note that the results for the two different societies are analogous. Considering the cases with $\tau = 10$, we can see that using perfect information (case 1) leads to membership functions that approximate the true distributions accurately. This can be seen in Table 8 from the error values, which are very close to zero for the *last 200* data points, for both the core center and the core width of the membership functions. This indicates that the proposed norm adaptation procedure can be employed in a real-time setting to accurately learn the interpretation of norms in a given society when perfect information is provided about how to interpret the situations. Similar results are obtained also for the inferred information (case 2). The results indicate that the membership functions are approximated effectively despite imperfect information, even though in some cases with a higher variability in the error.

In both cases, the core width error is more variable, because it is more affected by the data points collected and used for calculating the error: the fewer the data points, the stronger the influence that outliers have on the adaptation. This effect can be reduced by requiring more data points in order to perform the adaptation. We show this by evaluating the configuration with $\tau = 40$. In this case, the results show that the approach is less subject to outliers, i.e., compared to $\tau = 10$, the variability of all the errors is lower.

Figure 8 reports the trend for the errors in our simulations for one data set in the case of Society B (analogous results can be observed for all other cases). We note that the error quickly converges towards low values. As explained in

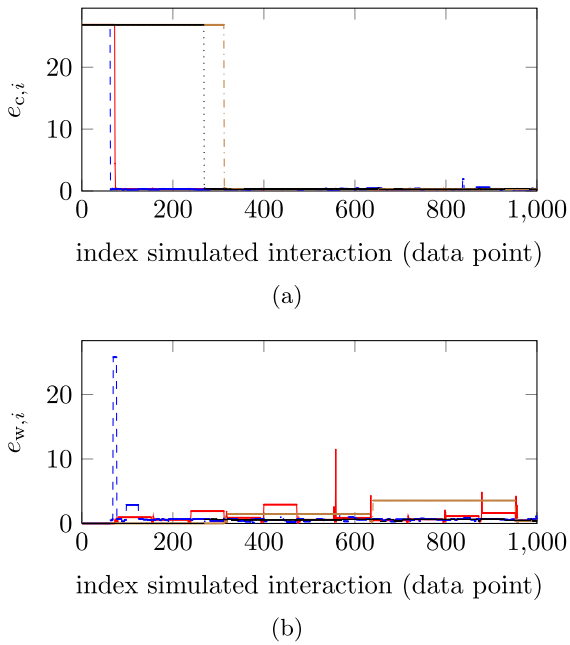


Fig. 8 Example for one data set for Society B showing the trend of the average RMSE w.r.t. the core center (**a**) and core width (**b**) after adding every data point for adaptation, for case 1 and $\tau = 10$ (see the red solid curve), for case 2 and $\tau = 10$ (see the blue dashed curve), for case 1 and $\tau = 40$ (see the brown dash-dotted curve), and for case 2 and $\tau = 40$ (see the black dotted curve)

Sect. 6.1.1, the considered data sets include 125 data points for each of the eight DIAMONDS situations. The data points are distributed evenly for each situation and according to the order reported in Table 7, i.e., the first, ninth, 17th, etc. data points concern the situation Duty, the second, 10th, 18th, etc. data points concern the situation Intellect, and so on for all the situations. Therefore, in case 1 (i.e., perfect information), the adaptation after collecting $\tau = 10$ data points is performed for the first time, for Duty, after the 73rd observed interaction (i.e., after the 10th data point has been collected for Duty). For $\tau = 40$, the first adaptation is performed after the 313th observed interaction (i.e., after the 40th data point is collected for Duty). The solid red and brown dash-dotted

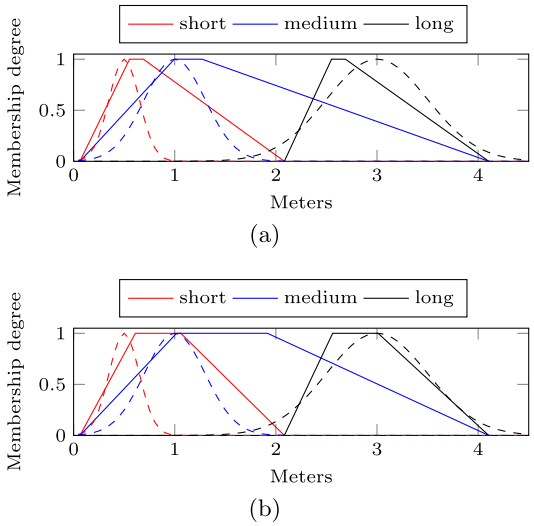


Fig. 9 Example for one data set for Society B and $\tau = 40$, showing the adapted membership functions for the fuzzy sets of variable *Distance* (whose real distributions from Table 6 are reported as dashed curves) after observing the 500*th* data point in the data set, for **a** case 1 and **b** case 2. Note that the figure only illustrates a snapshot of the membership functions after a certain observation. In the proposed continual learning approach, the values of the membership functions are in continuous change

curves in Fig. 8a clearly illustrate this concept and show that the error converges to low values as soon as the minimum number of data points, i.e., τ, for each situation variable has been collected. In case 2 (i.e., with imperfect information, shown by the blue dashed and black dotted curves Fig. 8), the first time that the adaptation is performed slightly varies due to some of the data points being wrongly classified as intended. However, the adaptation process still quickly converges as soon as the minimum number of data points has been collected.

After converging toward low values, the error oscillates around such values for the rest of the time. Oscillations follow from the continual learning approach that we proposed, which leverages the most recent data points to determine a norm adaptation and does not require to preserve the entire data set obtained so far.

The results reported above illustrates that the mechanism can effectively be applied in real time, since it does not require large amount of data, but converges as expected even when few data points are provided at a time. Moreover, the results illustrate that, if a robot endowed with such a mechanism is placed in a new society, it can quickly adapt its rules to align with the appropriate behaviors observed in that society. Figure 9 reports an example of snapshot of the membership functions of variable *Distance* for Society B after the error converged to low values.

7 Conclusions and Future Work

In this paper, we introduced a novel general-purpose and robot-agnostic control architecture, SONAR, standing for SOcial Norm Aware Robots. SONAR brings together various state-of-the-art technologies into an efficient control architecture for high-level automated decision making and adaptive norm-aware capabilities for social robots. By leveraging fuzzy logic and fuzzy inference, SONAR attributes social meanings to physical inputs received via the robot sensors in order to make a social interpretation of the situation where the robot operates. Based on the inferred social situation, SONAR determines appropriate, obliged, and prohibited actions, as well as modes of execution of those actions in line with the social norms and practices. Furthermore, through a continual learning approach, SONAR permits to learn social norms from data acquired during interactions with humans.

We evaluated the usability, perception, experience, and acceptance of a Nao robot steered via a Python implementation of SONAR through experiments of human-robot interactions. We considered scenarios where participants had a casual conversation with the robot, during which they performed five tasks (greeting, role playing game, discussing a personal issue, paying attention to an object, goodbye). The robot, steered by SONAR, interacted fully autonomously with the participants, by leveraging GPT-based large language models for natural language processing and generation, and normative reasoning for determining adequate and proactive behaviors based on the task being executed.

The results of our experiments indicate that our implementation of SONAR can be effectively and efficiently used in human-robot interactions (*RQ1.1*). Despite the exploratory nature of our study, the Nao-SONAR robot, leveraging social and norm awareness via SONAR, successfully completed about 80% of the tasks. The results also indicate that Nao-SONAR leads to more positive and enjoyable interactions with Nao, compared to using Nao-Chatbot, which leverages no explicit social and normative reasoning (*RQ1.2*). Nao-SONAR was perceived as more sociable, active, assertive, considerate, appropriate, reactive, proactive, and autonomous, compared to Nao-Chatbot. Communication with Nao-SONAR was experienced as easier than with Nao-Chatbot, even though both robots relied on the same language model.

We also investigated, via computer-based simulations, the extent to which SONAR can be used to learn social norms of a society. The results of our simulations indicate that the

proposed norm-adaptation mechanism can quickly learn new rules of behavior in a society, and requires little amount of data to adapt to new norms (*RQ2*).

Limitations The results from human-robot interactions indicate that further work is needed to improve the accuracy of our implementation, in particular concerning the detection of social cues and the intentions of humans. In the future, we intend to extend SONAR to refine its algorithms and parameters (e.g., for mining the intentions of humans with more precision from natural language). In addition, we intend to test better sensors (e.g., cameras with higher sensitivity and zooming capabilities) and image detection algorithms to improve the detection of social cues. Additionally, we intend to investigate efficient sensor fusion mechanisms that can further improve the gathering of the social cues and the interpretation of the context by using various (possibly non-homogeneous) data, such as temperature and light intensity.

Our experiments also indicate that some of the rules of behaviors and norms that we introduced were too simplistic and required more fine grained conditions to improve their accuracy. In real-world situations, accurate rules might require considering many cues and conditions (including the content of the speech, the voice tone, the body posture, etc.). Defining rules for all possible situations by hand, as currently done in this paper, is clearly not feasible in the general case and represents a limitation of our work. In future work, we aim to investigate how SONAR can autonomously elicit and learn norms, e.g., by employing learning techniques such as those discussed in [25, 27, 88, 110]. In this scenario, mechanisms for conflict resolution and filtering the rules, and for ensuring coherence of the rules should also be considered (e.g., [111, 112]).

Additionally, the quality of conversations in our experiments varied between participants. During the Introduction phase, participants could test the robot's understanding of their voice and adjust accordingly in order to ensure quality of the actual interactions. Despite this, in some interactions more than others, the detected speech was not always accurate. While we noted that, generally, the quality of interactions was affected by the accuracy of the speech recognition (with more interesting and natural conversations occurring when speech recognition was more accurate), we leave for future work an in-depth analysis of these aspects.

Finally, the evaluation of SONAR presented in this paper is not exhaustive and especially does not fully assess SONAR in comparison with other existing architectures. A systematic and formal comparative evaluation of the structure and behavior of the architecture is needed to adequately assess various properties such as consistency, completeness, and correctness of SONAR. Additionally, a systematic assessment of the scalability of SONAR in terms of the number of agents and the computational load that could be handled by these agents in real time, needs to be conducted. Further experiments concerning the norm-adaptation algorithms are required to assess their effectiveness in learning and adapting to personal norms, in addition to the societal ones. We also intend to introduce support for considering, during norm-adaptation, larger variations in the multiple social interpretations that can be attributed to a given situation.

Future research directions The participants in our study that involved a Nao humanoid robot indicated no preference for robot companions to have human-like appearances. These findings that are in line with uncanny valley theory [113] (the hypothesis that highly realistic humanoid robots will risk eliciting eerie feelings in people), deserve further investigation. In future research, we intend to integrate our implementation of SONAR with various humanoid and non-humanoid robots to explore whether or not the naturalness and acceptance of SONAR-based social companions is affected by the uncanny valley effect.

Similarly, we intend to investigate whether user perceptions vary when the robot considers different norms and how various types of norms influence user perception. This exploration could pave the way for intriguing studies in human-robot interaction (HRI) encompassing cultural dimensions. While our initial experiments offer preliminary insights in this direction-comparing a system incorporating various norms (Nao-SONAR) with a norm-agnostic system (Nao-Chatbot)-further investigation remains a critical aspect of our future research agenda. We hope that the promising outcomes presented herein will also inspire and facilitate other fellow researchers to embark on similar studies leveraging SONAR.

Our future work also includes delving into the effectiveness of SONAR in socially assistive contexts, where Socially Assistive Robots [114, 115] are increasingly used to implement, for example, robot-mediated therapeutic interventions in autism spectrum disorder [116, 117] or in dementia care [118]. This future work will entail encoding and learning the social norms that are tailored to the therapeutic domain and to the individual patient. These norms can be designed based on established guidelines [119], common practices in accepted therapeutic interventions [118], and personalized indications given by the caregivers and available medical knowledge about the patient (similarly to [25]). Additionally, future work should investigate how to automatically elicit and learn (e.g., as in [25, 27, 88, 110]) personal norms that characterize individual patient preferences, for example from dialogues and interactions with the patients and caregivers.

An interesting direction for future work is the integration of safety rules (e.g., safety zones) in our architecture [120, 121]. More particularly, we hypothesize that safety rules could be represented via norms encoded via fuzzy rules, and that normative reasoning could ensure safe human-robot collaboration in a shared workspace.

Finally, we believe that a hybrid architecture like SONAR, which combines symbolic and sub-symbolic reasoning and learning, could support several important approaches for robots in human-centered environments [122] and for hybrid intelligence systems [3, 99] beyond human-robot systems, which we intend to explore in future work. These include a computational theory of mind [123], multi-agent communication, and human-AI (norm-based) explainability approaches [30, 124].

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Davide Dell'Anna, and Anahita Jamshidnejad. The first draft of the manuscript was written by Davide Dell'Anna and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the NWO - Open Competition Domain Science XS project "Human-like norm-aware cognitive robots for autonomous interactions with humans" (OCENW.XS21.3.106), which has been financed by the Netherlands Organisation for Scientific Research (NWO).

Data Availability All the data generated and/or analyzed during the current study are available in the Zenodo repository https://doi.org/10.5281/zenodo.10719808. The videos of the human-robot interactions are available upon request via https://doi.org/10.4121/50c7a19c-fc0e-4ef3-b35a-dd23bf08470d.

Code Availability The source code generated and/or analyzed during the current study are available in the Zenodo repositories https://doi.org/10.5281/zenodo.10719808 (for the SONAR source code), and https://doi.org/10.5281/zenodo.7979416 (for the Python MQTT Nao Interface used to interface SONAR with Nao).

Declarations

Conflict of interest The authors declare they have no Conflict of interest.

Ethics Approval This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Human Research Ethics Committee of TU Delft under Approval No. 2584.

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Consent for Publication The authors affirm that human research participants provided informed consent for publication of the text, pictures and videos published in the article. Additional informed consent was obtained from all individual participants for whom identifying information is included in this article, which include the video recording reported in the supplementary material and made available upon request, and the images in Fig. 1.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Góngora Alonso S, Hamrioui S, de la Torre Díez I, Motta Cruz E, López-Coronado M, Franco M (2019) Social robots for people with aging and dementia: a systematic review of literature. Telemed e-Health 25(7):533–540
2. Ferrara E, Varol O, Davis CA, Menczer F, Flammini A (2016) The rise of social bots. Commun ACM 59(7):96–104
3. Akata Z, Balliet D, de Rijke M, Dignum F, Dignum V, Eiben G et al (2020) A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer 53(8):18–28
4. Bicchieri C (2005) The grammar of society: the nature and dynamics of social norms. Cambridge University Press, Cambridge
5. Lewis D (2008) Convention: a philosophical study. Wiley, Hoboken
6. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. First edition.," 2019
7. Brinck I, Balkenius C, Johansson B (2016) Making place for social norms in the design of human-robot interaction. In: What social robots can and should do—proceedings of robophilosophy 2016/TRANSOR 2016, Aarhus, Denmark, October 17–21, 2016. vol. 290 of Frontiers in Artificial Intelligence and Applications. IOS Press, pp 303–312
8. Avelino J, Garcia-Marques L, Ventura R, Bernardino A (2021) Break the ice: a survey on socially aware engagement for human–robot first encounters. Int J Soc Robot 13(8):1851–1877
9. Rasouli S, Gupta G, Nilsen E, Dautenhahn K (2022) Potential applications of social robots in robot-assisted interventions for social anxiety. Int J Soc Robot 1–32
10. Wasik A, Tomic S, Saffiotti A, Pecora F, Martinoli A, Lima PU (2018) Towards norm realization in institutions mediating human–robot societies. In: 2018 IEEE/RSJ international conference on intelligent robots and systems, IROS 2018, Madrid, Spain, October 1–5, 2018. IEEE, pp 297–304
11. Bruno B, Mastrogiovanni F, Pecora F, Sgorbissa A, Saffiotti A (2017) A framework for culture-aware robots based on fuzzy logic. In: 2017 IEEE international conference on fuzzy systems, FUZZ-IEEE 2017, Naples, Italy, July 9–12, 2017. IEEE, pp 1–6
12. Cranefield S, Dignum F (2019) Incorporating social practices in BDI agent systems. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, AAMAS '19, Montreal, QC, Canada, May 13–17, 2019. International Foundation for Autonomous Agents and Multiagent Systems, pp 1901–1903
13. Mokkapati S (2021) Implementation of social practices on the pepper robot in the elderly care domain: AI planning with social practices [Master's Thesis, Umeå University]. Digitala Vetenskapliga Arkivet. https://www.diva-portal.org/smash/get/diva2:1605036/FULLTEXT01.pdf
14. Von Wright GH (1951) Deontic logic. Mind 60(237):1–15
15. Conte R, Castelfranchi C, Dignum F (1998) Autonomous norm acceptance. In: Proceedings of the 5th international workshop on agent theories, architectures, and languages, ATAL 1998. pp 99–112

16. Dignum F, Kuiper R (1997) Combining dynamic deontic logic and temporal logic for the specification of deadlines. In: Proceedings of the 30th annual Hawaii international conference on system sciences, HICSS-30, pp 336–346
17. Wieringa R, Meyer JJC (1993) Actors, actions, and initiative in normative system specification. Ann Math Artif Intell 7(1–4):289–346
18. Kripke S (2007) Semantical considerations of the modal logic. Stud Philos 1
19. Gamblin S, Niveau A, Bouzid M (2022) A symbolic representation for probabilistic dynamic epistemic logic. In: 21st International conference on autonomous agents and multiagent systems, AAMAS 2022, Auckland, New Zealand, May 9–13, 2022. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), pp 445–453
20. Cranefield S, Savarimuthu BTR (2021) Normative multi-agent systems and human-robot interaction. In: Workshop on robot behavior adaptation to human social norms (TSAR), pp 1–3
21. Bruno B, Chong NY, Kamide H, Kanoria S, Lee J, Lim Y, et al (2017) Paving the way for culturally competent robots: a position paper. In: 26th IEEE International symposium on robot and human interactive communication, RO-MAN 2017, Lisbon, Portugal, August 28–September 1, 2017. IEEE, pp 553–560
22. Bruno B, Chong NY, Kamide H, Kanoria S, Lee J, Lim Y, et al (2017) The CARESSES EU-Japan project: making assistive robots culturally competent. In: Ambient assisted living—Italian Forum 2017, Eighth Italian on Ambient Assisted Living Forum, ForItAAL 2017, 14–15 June, 2017, Genoa, Italy. vol. 540 of Lecture Notes in Electrical Engineering. Springer, pp 151–169
23. Bai Y, Wang D (2006) Fundamentals of fuzzy logic control—fuzzy sets, fuzzy rules and defuzzifications. In: Bai Y, Zhuang H, Wang D (eds) Advanced fuzzy logic technologies in industrial applications. Advances in industrial control. Springer, London, pp 17–36
24. Mobahi H, Ansari S (2003) Fuzzy perception, emotion and expression for interactive robots. In: Proceedings of the IEEE international conference on systems, man & cybernetics: Washington, DC, USA, 5–8 October 2003. IEEE, pp 3918–3923
25. Dell'Anna D, Jamshidnejad A (2022) Evolving fuzzy logic systems for creative personalized socially assistive robots. Eng Appl Artif Intell 114:105064
26. Vitiello A, Acampora G, Staffa M, Siciliano B, Rossi S (2017) A neuro-fuzzy-Bayesian approach for the adaptive control of robot proxemics behavior. In: Proceedings of the 2017 IEEE international conference on fuzzy systems, FUZZ-IEEE 2017, Naples, Italy, July 9–12, 2017. IEEE, pp 1–6
27. Savarimuthu BTR, Cranefield S (2011) Norm creation, spreading and emergence: a survey of simulation models of norms in multi-agent systems. Multiagent Grid Syst 7(1):21–54
28. Dell'Anna D, Dastani M, Dalpiaz F (2020) Runtime revision of sanctions in normative multi-agent systems. Auton Agents Multiagent Syst 34(2):43
29. Bratman M et al (1987) Intention, plans, and practical reason, vol 10. Harvard University Press, Cambridge
30. Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38
31. Winikoff M, Sidorenko G, Dignum V, Dignum F (2021) Why bad coffee? Explaining BDI agent behaviour with valuings. Artif Intell 300:103554
32. Gudwin RR, Gomide FAC, Pedrycz W (1998) Context adaptation in fuzzy processing and genetic algorithms. Int J Intell Syst 13(10–11):929–948
33. Robotics S.: NAO the humanoid and programmable robot. https://www.softbankrobotics.com/emea/en/nao. Accessed on 07/05/2022
34. Dell'Anna D, Jamshidnejad A.: SONAR: an adaptive control architecture for SOcial norm aware robots—code and supplementary material. Zenodo. Available from: https://doi.org/10.5281/zenodo.10719808
35. Dell'Anna D, Jamshidnejad A.: Dataset: video recordings of human-robot interactions with a Nao robot controlled via the SONAR adaptive control architecture for social norm aware robots. 4TU.ResearchData. https://doi.org/10.4121/50c7a19c-fc0e-4ef3-b35a-dd23bf08470d
36. Searle JR, Willis YS, et al (1995) The construction of social reality. Simon and Schuster
37. Elster J (1989) The cement of society: a survey of social order. Cambridge University Press, Cambridge
38. Gibbs JP (1965) Norms: the problem of definition and classification. Am J Sociol 70(5):586–594
39. Kanger S.: New foundations for ethical theory. In: Deontic logic: introductory and systematic readings. Ed. Hilpinen, R. Reidel Publishing Company
40. Epstein JM (2001) Learning to be thoughtless: social norms and individual computation. Comput Econ 18(1):9–24
41. North DC (1990) Institutions and a transaction-cost theory of exchange. Perspect Posit Polit Econ 182:191
42. Castelfranchi C, Dignum F, Jonker CM, Treur J (1999) Deliberative normative agents: principles and architecture. In: Proceedings of the 6th international workshop on agent theories, architectures, and languages, ATAL 1999. pp 364–378
43. Rato D, Prada R (2021) Towards social identity in socio-cognitive agents. Sustainability 13(20):11390
44. Dignum F, Dignum V, Prada R, Jonker CM (2015) A conceptual architecture for social deliberation in multi-agent organizations. Multiagent Grid Syst 11(3):147–166
45. Dignum F, Dignum V (2020) How to center AI on humans. In: Proceedings of the first international workshop on new foundations for human-centered AI (NeHuAI) co-located with 24th European conference on artificial intelligence (ECAI 2020), Santiago de Compostella, Spain, September 4, 2020. vol. 2659 of CEUR workshop proceedings. CEUR-WS.org. pp 59–62
46. Castro VF, Hakli R, Clodic A (2020) What does it take to be a social agent? In: Culturally sustainable social robotics—proceedings of Robophilosophy 2020, Virtual Event, 2020. vol. 335 of Frontiers in Artificial Intelligence and Applications. IOS Press, pp 540–549
47. Broersen JM, Dastani M, van der Torre LWN (2005) Beliefs, obligations, intentions, and desires as components in an agent architecture. Int J Intell Syst 20(9):893–919
48. Broersen JM, Dastani M, Hulstijn J, Huang Z, van der Torre LWN (2001) The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In: Proceedings of the fifth international conference on autonomous agents, AGENTS 2001. pp 9–16
49. Bhattacharya P, de Mooij AJ, Dell'Anna D, Dastani M, Logan B, Swarup S (2021) PanSim + Sim-2APL: a framework for large-scale distributed simulation with complex agents. In: Engineering multi-agent systems—9th international workshop, EMAS 2021, Virtual Event, May 3–4, 2021, Revised selected papers. vol. 13190 of Lecture Notes in Computer Science. Springer, pp 1–21
50. de Mooij J, Bhattacharya P, Dell'Anna D, Dastani M, Logan B, Swarup S (2023) A framework for modeling human behavior in large-scale agent-based epidemic simulations. Simulation 99(12):1183–1211
51. Bratman ME, Israel DJ, Pollack ME (1988) Plans and resource-bounded practical reasoning. Comput Intell 4:349–355
52. Georgeff MP, Lansky AL (1987) Reactive reasoning and planning. In: Proceedings of the 6th national conference on artificial intelligence, AAAI 1987, pp 677–682

53. Rao AS (1996) AgentSpeak(L): BDI agents speak out in a logical computable language. In: Proceedings of the 7th European workshop on modelling autonomous agents in a multi-agent world, MAAMAW 1996, pp 42–55
54. Bordini RH, Hübner JF, Vieira R (2005) Jason and the golden fleece of agent-oriented programming. In: Multi-agent programming: languages, platforms and applications. pp 3–37
55. Pokahr A, Braubach L, Lamersdorf W (2005) Jadex: A BDI reasoning engine. In: Multi-agent programming: languages, platforms and applications, pp 149–174
56. Shoham Y (1993) Agent-oriented programming. Artif Intell 60(1):51–92
57. Castelfranchi C (1999) Prescribed mental attitudes in goal-adoption and norm-adoption. Artif Intell Law 7(1):37–50
58. Dignum F, Morley DN, Sonenberg L, Cavedon L (2000) Towards socially sophisticated BDI agents. In: Proceedings of the fourth international conference on multi-agent systems, ICMAS 2000, pp 111–118
59. Kc U, Chodorowski J (2019) A case study of adding proactivity in indoor social robots using belief-desire-intention (BDI) model. Biomimetics 4(4):74
60. Ziafati P, Dastani M, Meyer JC, van der Torre LWN (2012) Agent programming languages requirements for programming autonomous robots. In: Programming multi-agent systems—10th international workshop, ProMAS 2012, Valencia, Spain, June 5, 2012, Revised Selected Papers. vol. 7837 of Lecture Notes in Computer Science. Springer, pp 35–53
61. Wesz RB (2015) Integrating robot control into the Agentspeak (L) programming language [Master's thesis]. Pontifícia Universidade Católica do Rio Grande do Sul
62. Ricci A, Viroli M, Omicini A (2006) CArtAgO: A framework for prototyping artifact-based environments in MAS. In: International workshop on environments for multi-agent systems. Springer, pp 67–86
63. Ribino P, Bonomolo M, Lodato C, Vitale G (2021) A humanoid social robot based approach for indoor environment quality monitoring and well-being improvement. Int J Soc Robot 13(2):277–296
64. Correia F, Campos J, Melo FS, Paiva A (2023) Robotic gaze responsiveness in multiparty teamwork. Int J Soc Robot 15(1):27–36
65. Filippini C, Merla A (2023) Systematic review of affective computing techniques for infant robot interaction. Int J Soc Robot 15(3):393–409
66. Chen C, Jia X (2023) Effects of head shape, facial features, camera, and gender on the perceptions of rendered robot faces. Int J Soc Robot 15(1):71–84
67. Higashino K, Kimoto M, Iio T, Shimohara K, Shiomi M (2023) Is politeness better than impoliteness? Comparisons of Robot's encouragement effects toward performance, moods, and propagation. Int J Soc Robot 1–13
68. Kola I, Jonker CM, van Riemsdijk MB (2019) Who's that? Social situation awareness for behaviour support agents—a feasibility study. In: Engineering multi-agent systems—7th international workshop, EMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers. vol. 12058 of Lecture Notes in Computer Science. Springer, pp 127–151
69. Rauthmann JF, Gallardo-Pujol D, Guillaume EM, Todd E, Nave CS, Sherman RA et al (2014) The situational eight DIAMONDS: a taxonomy of major dimensions of situation characteristics. J Pers Soc Psychol 107(4):677
70. Repiso E, Garrell A, Sanfeliu A (2022) Adaptive social planner to accompany people in real-life dynamic environments. Int J Soc Robot 1–33
71. Johanson DL, Ahn HS, Broadbent E (2021) Improving interactions with healthcare robots: a review of communication behaviours in social and healthcare contexts. Int J Soc Robot 13(8):1835–1850
72. Carlucci FM, Nardi L, Iocchi L, Nardi D (2015) Explicit representation of social norms for social robots. In: 2015 IEEE/RSJ international conference on intelligent robots and systems, IROS 2015, Hamburg, Germany, September 28–October 2, 2015. IEEE, pp 4191–4196
73. Long SA, Esterline AC (2000) Fuzzy BDI architecture for social agents. In: Proceedings of the IEEE SoutheastCon 2000. Preparing for The New Millennium (Cat. No. 00CH37105). IEEE, pp 68–74
74. Cruz A, dos Santos AV, Santiago RH, Bedregal B (2021) A fuzzy semantic for BDI logic. Fuzzy Inf Eng 13(2):139–153
75. Farias GP, Dimuro GP, da Rocha Costa AC (2010) BDI agents with fuzzy perception for simulating decision making in environments with imperfect information. In: Proceedings of the multi-agent logics, languages, and organisations federated workshops (MALLOW 2010), Lyon, France, August 30–September 2, 2010. vol. 627 of CEUR workshop proceedings. CEUR-WS.org
76. Cuesta F, Ollero A, Arrue BC, Braunstingl R (2003) Intelligent control of nonholonomic mobile robots with fuzzy perception. Fuzzy Sets Syst 134(1):47–64
77. Mobahi H, Ansari S (2003) Fuzzy perception, emotion and expression for interactive robots. In: Proceedings of the IEEE international conference on systems, man & cybernetics: Washington, DC, USA, 5–8 October 2003. IEEE, pp 3918–3923
78. Dimuro G, Santos A, Bedregal G, Costa A, Lopes L, Lau N, et al (2009) Fuzzy evaluation of social exchanges between personality-based agents. In: New trends in artificial intelligence, proceedings of the 14th Portuguese conference on artificial intelligence. APIA, Aveiro, pp 451–462
79. Elkosantini S, Gien D (2007) Human behavior and social network simulation: fuzzy sets/logic and agents-based approach. In: Ades MJ (eds) Proceedings of the 2007 spring simulation multiconference, SpringSim 2007, Norfolk, Virginia, USA, March 25–29, 2007, Volume 2. SCS/ACM, pp 102–109
80. Hassan S, Salgado M, Pavón J (2008) Friends forever: social relationships with a fuzzy agent-based model. In: Hybrid artificial intelligence systems, third international workshop, HAIS 2008, Burgos, Spain, September 24–26, 2008. Proceedings. vol. 5271 of Lecture Notes in Computer Science. Springer, pp 523–532
81. Fort H, Pérez N (2005) The fate of spatial dilemmas with different fuzzy measures of success. J Artif Soc Soc Simul 8(3)
82. Boella G, van der Torre LWN (2004) Regulative and constitutive norms in normative multiagent systems. In: Principles of knowledge representation and reasoning: proceedings of the ninth international conference (KR2004), Whistler, Canada, June 2–5, 2004. AAAI Press, pp 255–266
83. Wooldridge M (2009) An introduction to multiagent systems. Wiley, Hoboken
84. Palanca J, Terrasa A, Julián V, Carrascosa C (2020) SPADE 3: supporting the new generation of multi-agent systems. IEEE Access 8:182537–182549
85. Saint-Andre, P (2011) Extensible Messaging and Presence Protocol (XMPP): Core; Technical Report; RFC Editor, Internet Engineering Task Force: Fremont, CA, USA
86. O'Brien PD, Nicol RC (1998) FIPA-towards a standard for software agents. BT Technol J 16:51–59
87. International Organization for Standardization.: ISO/IEC 20922:2016(en), Information technology—message queuing telemetry transport (MQTT) v3.1.1. https://www.iso.org/obp/ui/#iso:std:iso-iec:20922:ed-1:v1:en. Accessed on 07/05/2022
88. Dell'Anna D, Alechina N, Dalpiaz F, Dastani M, Logan B (2022) Data-driven revision of conditional norms in multi-agent systems. J Artif Intell Res 75:1549–1593

89. Botta A, Lazzerini B, Marcelloni F, Stefanescu DC (2009) Context adaptation of fuzzy systems through a multi-objective evolutionary approach based on a novel interpretability index. Soft Comput 13(5):437–449
90. Sadollah A (2018) Introductory chapter: which membership function is appropriate in fuzzy system? In: Fuzzy logic based in optimization methods and control systems and its applications. IntechOpen
91. International Organization for Standardization.: ISO/IEC 25022:2016 systems and software engineering—systems and software quality requirements and evaluation (SQuaRE)–measurement of quality in use https://www.iso.org/standard/35746.html. Accessed on 22/02/2024
92. Nomura T, Kanda T, Suzuki T (2006) Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. AI Soc 20(2):138–150
93. Woods S, Dautenhahn K, Kaouri C, te Boekhorst R, Koay KL (2005) Is this robot like me? Links between human and robot personality traits. In: 5th IEEE-RAS international conference on humanoid robots, humanoids 2005, Tsukuba, Japan, December 5–7, 2005. IEEE, pp 375–380
94. Sunstein CR (1996) Social norms and social roles. Colum L Rev 96:903
95. Commission E (2019) Directorate-general for communications networks C: Technology. Ethics guidelines for trustworthy AI. Publications Office
96. Parliament E (2023) EU artificial intelligence act. European Parliament
97. Fujii G, Hamada K, Ishikawa F, Masuda S, Matsuya M, Myojin T et al (2020) Guidelines for quality assurance of machine learning-based artificial intelligence. Int J Softw Eng Knowl Eng 30(11&12):1589–1606
98. Contreras I, Vehi J (2018) Artificial intelligence for diabetes management and decision support: literature review. J Med Int Res 20(5):e10775
99. Dell'Anna D, Murukannaiah PK, Dudzik B, Grossi D, Jonker CM, Oertel C, et al (2024) Toward a quality model for hybrid intelligence teams. In: Proceedings of the 23rd international conference on autonomous agents and multiagent systems, AAMAS 2024. ACM, pp 463–464
100. Weiss A, Bernhaupt R, Lankes M, Tscheligi M (2009) The USUS evaluation framework for human–robot interaction. In: AISB2009: Proceedings of the symposium on new frontiers in human–robot interaction. vol. 4. pp 11–26
101. Dell'Anna D.: MQTT-nao-interface: a Python 2.7 interface for the Nao robot based on MQTT. Zenodo. https://doi.org/10.5281/zenodo.7979416
102. Zhang Y, Sun S, Galley M, Chen Y, Brockett C, Gao X, et al (2020) DIALOGPT: large-scale generative pre-training for conversational response generation. In: Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations, ACL 2020, Online, July 5–10, 2020. Association for Computational Linguistics, pp 270–278
103. Jung M, Lazaro MJS, Yun MH (2021) Evaluation of methodologies and measures on the usability of social robots: a systematic review. Appl Sci 11(4):1388
104. Miller RB (1968) Response time in man-computer conversational transactions. In: American federation of information processing societies: proceedings of the AFIPS '68 fall joint computer conference, December 9–11, 1968, San Francisco, California, USA—Part I. vol. 33 of AFIPS conference proceedings. AFIPS/ACM/Thomson Book Company, Washington DC, pp 267–277
105. Starner T (2001) The challenges of wearable computing: Part 2. IEEE Micro 21(4):54–67
106. Shiwa T, Kanda T, Imai M, Ishiguro H, Hagita N (2009) How quickly should a communication robot respond? Delaying strategies and habituation effects. Int J Soc Robot 1(2):141–155
107. Pallant J (2011) Survival manual. A step by step guide to data analysis using SPSS. 4:4
108. Vinciarelli A, Pantic M, Bourlard H (2009) Social signal processing: survey of an emerging domain. Image Vis Comput 27(12):1743–1759
109. NCEH, CDC.: What noises cause hearing loss?. https://www.cdc.gov/nceh/hearing_loss/what_noises_cause_hearing_loss.html. Accessed on 01/23/2023
110. Wang LX, Mendel JM (1992) Generating fuzzy rules by learning from examples. IEEE Trans Syst Man Cybern 22(6):1414–1427
111. Dubois D, Prade H, Ughetto L (1997) Checking the coherence and redundancy of fuzzy knowledge bases. IEEE Trans Fuzzy Syst 5(3):398–417
112. Vasconcelos WW, Kollingbaum MJ, Norman TJ (2009) Normative conflict resolution in multi-agent systems. Auton Agent Multiagent Syst 19:124–152
113. Mori M, MacDorman KF, Kageki N (2012) The uncanny valley [from the Field]. IEEE Robotics Autom Mag 19(2):98–100
114. Van Wynsberghe A (2016) Healthcare robots: ethics, design and implementation. Routledge, London
115. Syriopoulou-Delli CK, Gkiolnta E (2022) Review of assistive technology in the training of children with autism spectrum disorders. Int J Dev Disabil 68(2):73–85
116. Nawaz R, Ali S (2022) Introducing therapeutic robotics for autism. Emerald Publishing Limited
117. Ali S, Mehmood F, Dancey D, Ayaz Y, Khan MJ, Naseer N et al (2019) An adaptive multi-robot therapy for improving joint attention and imitation of ASD children. IEEE Access 7:81808–81825
118. Van Mierlo L, Van der Roest H, Meiland F, Dröes R (2010) Personalized dementia care: proven effectiveness of psychosocial interventions in subgroups. Ageing Res Rev 9(2):163–183
119. WHO: Integrated care for older people. https://www.who.int/publications/i/item/WHO-FWC-ALC-19.1
120. Galin R, Meshcheryakov RV (2019) Review on human–robot interaction during collaboration in a shared workspace. In: Interactive collaborative robotics—4th international conference, ICR 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings. vol. 11659 of Lecture Notes in Computer Science. Springer, pp 63–74
121. Augustsson S, Christiernin LG, Bolmsjö G (2014) Human and robot interaction based on safety zones in a shared work environment. In: ACM/IEEE international conference on human-robot interaction, HRI'14, Bielefeld, Germany, March 3–6, 2014. ACM, pp 118–119
122. Gomoll AS, Sabanovic S, Tolar E, Hmelo-Silver CE, Francisco MR, Lawlor OS (2018) Between the social and the technical: negotiation of human-centered robotics design in a middle school classroom. Int J Soc Robot 10(3):309–324
123. Carruthers P, Smith PK (1996) Theories of theories of mind. Cambridge University Press, Cambridge
124. Umbrello S, Yampolskiy RV (2022) Designing AI for explainability and verifiability: a value sensitive design approach to avoid artificial stupidity in autonomous vehicles. Int J Soc Robot 14(2):313–322

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.