electronics

MDPI

Article

Redefining User Expectations: The Impact of Adjustable Social Autonomy in Human–Robot Interaction

Filippo Cantucci [1,*], Rino Falcone [1] and Marco Marini [2]

[1] Institute of Cognitive Sciences and Technologies, National Research Council of Italy (ISTC-CNR), 00185 Rome, Italy; rino.falcone@istc.cnr.it
[2] IMT School for Advanced Studies Lucca, 55100 Lucca, Italy; marco.marini@imtlucca.it
* Correspondence: filippo.cantucci@istc.cnr.it

Abstract: To promote the acceptance of robots in society, it is crucial to design systems exhibiting adaptive behavior. This is particularly needed in various social domains (e.g., cultural heritage, healthcare, education). Despite significant advancements in adaptability within Human-Robot Interaction and Social Robotics, research in these fields has overlooked the essential task of analyzing the robot's cognitive processes and their implications for intelligent interaction (e.g., adaptive behavior, personalization). This study investigates human users' satisfaction when interacting with a robot whose decision-making process is guided by a computational cognitive model integrating the principles of adjustable social autonomy. We designed a within-subjects experimental study in the domain of Cultural Heritage, where users (e.g., museum visitors) interacted with the humanoid robot Nao. The robot's task was to provide the user with a museum exhibition to visit. The robot adopted the delegated task by exerting some degree of discretion, which required different levels of autonomy in the task adoption, relying on its capability to have a theory of mind. The results indicated that as the robot's level of autonomy in task adoption increased, user satisfaction with the robot decreased, whereas their satisfaction with the tour itself improved. Results highlight the potential of adjustable social autonomy as a paradigm for developing autonomous adaptive social robots that can improve user experiences in multiple HRI real domains.

Keywords: human–robot interaction; adaptive social robots; behavior adaptability; cognitive systems; adjustable social autonomy; theory of mind

Citation: Cantucci, F.; Falcone, R.; Marini, M. Redefining User Expectations: The Impact of Adjustable Social Autonomy in Human–Robot Interaction. *Electronics* **2024**, *13*, 127. https://doi.org/10.3390/electronics13010127

Academic Editors: Rania Hodhod, Mohammad Jafari and Stefanos Kollias

Received: 20 October 2023
Revised: 12 December 2023
Accepted: 18 December 2023
Published: 28 December 2023

1. Introduction

Human interaction with autonomous robots has become increasingly ubiquitous [1]. To foster the acceptance of robots in society, it is crucial to deploy systems capable of adapting their behavior both to the environment and the needs of interacting users [2]. This is challenging across multiple social domains, including cultural heritage [3,4], elderly assistance [5], and tourism [6], among others. Despite significant advancements in adaptivity within Human–Robot Interaction (HRI) [7] and Social Robotics [1], research overlooked the critical role of analyzing the robot's cognitive processes and their implications for intelligent interaction, such as adaptive behavior [7], personalization, and so on. Therefore, it is essential to explore alternative interaction methodologies that facilitate the design of robots aligning with human expectations, capabilities, and limitations within specific contexts. The starting point is to identify the priority needs that humans expect to satisfy during effective interactions with their peers and, consequently, to shift our focus to HRI scenarios. Humans need to:

1. Pursue their own goals, i.e., the state of affairs the user wants to achieve;
2. Consider broader interests or active goals, i.e., the state of affairs that the user, with a specific profile and set of mental states, could be interested in or has already planned to achieve.

Point 1 is to achieve goals in the context of the ongoing interaction. In HRI, this implies building robots capable of understanding these goals and selecting an appropriate strategy (e.g., plan) to accomplish them. The level of user satisfaction will depend on the plan the robot intends to execute and the boundary conditions it employs. Point 2 arises whenever humans have interests and goals that extend beyond the current interaction. To be an effective collaborator, a robot should comprehend these implicit interests and goals and work towards their achievement or protection. For instance, a task delegated by a human could be part of a much more complex goal, or the same task might conflict with other goals that the human has not consciously considered. In both scenarios, the robot's role goes beyond simple task execution and requires a more complex assessment of the human's mental states (e.g., intentions, beliefs, motivations, interests); the robot must select a plan that aligns with those mental states. This can imply a mismatch between the task delegated and the one adopted.

Imagine a situation where a human museum visitor is equipped with sensors capable of collecting physiological data linked to the visitor's emotional states. The visitor delegates the robot to plan a visit to the room where a renowned Picasso painting is displayed. However, instead of merely adhering to the visitor's explicit request (user's goal), the robot decides to propose alternative tour options. For example, based on the visitor's profile and her passion for the Italian Renaissance, the robot might recommend visiting Michelangelo's room, as Picasso's painting is located in a crowded area with long wait times. Alternatively, if the robot observes a drop in the visitor's blood sugar levels, it might recommend a brief stop at the museum's bar to address the potential health concern. In more critical situations, such as detecting an abnormal heart rate, the robot may even initiate a call to a doctor or an ambulance. In these alternative scenarios, the robot considers the user's goals and interests, which the user may not be consciously aware of or explicitly acknowledge. These considerations may not always align with the task initially assigned to the robot.

The present study investigates human user satisfaction when interacting with a robot whose decision-making process is guided by a computational cognitive model [3] integrating the principles of adjustable social autonomy [8]. We designed a within-subjects experiment in cultural heritage, where museum visitors interacted with the humanoid robot Nao [9]. The robot's task was to provide the user with a museum exhibition to visit. The rest of the paper is organized as follows: Section 2 provides a theoretical background on the approach exploited in the present work; Section 3 gives an overview of the state of the art in adaptivity in HRI and in the specific domain of museums; Section 4 presents the experimental methods applied in our experiment with the Nao robot; in Section 5, we describe the statystical analysis exploited for analyzing the results; Section 6 presents the experimental findings; in Section 7 we discuss the results obtained; Section 8 is dedicated to conclusions and future work.

2. Background

Adjustable social autonomy [8] models the complex scenario in which a cognitive agent (an agent with its own beliefs and goals) must decide if and how to delegate or adopt a task for another agent within a given context and how much autonomy is necessary for that specific task. Falcone and Castelfranchi [8] extensively explored how any collaborative scenario involving intelligent agents (human or artificial) inherently implies the two basic complex attitudes of task delegation and adoption. The task delegation and adoption model [10] delineates the process in which an agent X (the client) delegates to an agent Y (the contractor), in a context C, to execute a task τ and to realize the result p (state of world), that includes or corresponds to X's goal $Goal_x(g) = g_x$. The task τ, the object of delegation, can be referred to as an action α and/or to its resulting world's state p. By means τ, we will refer to the action/state of world pair $\tau = (\alpha, p)$.

According to [8], the contractor can adopt the task τ at different levels of autonomy:

- *Literal help*: the contractor adopts exactly what has been delegated by the client;
- *Sub help*: the contractor satisfies just a sub-goal of the delegated task;

- *Over help*: the contractor goes beyond what has been delegated by the client without changing the client's plan;
- *Critical help*: the contractor satisfies the relevant results of the requested plan/action, but modifies that plan/action;
- *Critical Sub help*: the contractor realizes a sub help and in addition modifies the plan/action;
- *Critical Over help*: the contractor realizes an over help and in addition modifies the plan/action;
- *Hyper-critical help*: the contractor adopts goals or interests of the client that the client itself did not take into account (at least, in that specific interaction with the contractor): by doing so, the contractor neither performs the action/plan nor satisfies the results that were delegated.

In our HRI scenario, client X is the user, whereas contractor Y is the robot. The theory of adjustable social autonomy [8] represents the theoretical background underlying the design of the present study. Additionally, we leveraged a crucial theoretical concept applied in the field of HRI known as the theory of mind (ToM).

Theory of mind [11,12] can be defined as the ability of an agent, human or artificial, to ascribe to other agents specific mental states, and to take them into account for making decisions. Modeling other agents' mental states is one of the most important abilities learned by humans when they engage in cooperative interactions. Providing robots with the capability to build complex models of the interlocutors' mental states and to adapt their decisions based on these models represents an important aspect in promoting intelligent collaboration based on adjustable social autonomy.

The theoretical models mentioned earlier have been applied to create a cognitive computational model [3,13], defining the behavior of the robot in this experiment. Without going into detail, the proposed cognitive model defines a task-oriented, belief–desire–intention (BDI) [14,15] robot that performs effective reasoning within dynamic interaction with other cognitive agents, typically humans (Figure 1). The robot adopts a delegated task τ and exerts some degree of discretion, which corresponds to various levels of autonomy in τ adoption. The capability to expand the decision and action space concerning τ shows the adaptability to the user's mental states as a criterion. In fact, this adaptability is possible thanks to the robot's capability to have a ToM, enabling it to attribute mental states beyond those explicitly declared (we defined them as active goals). The user's model is built following a human profiling phase, during which the robot extracts relevant user features via verbal and non-verbal communication. After that, the robot elicits an internal negotiation process in which it mediates τ adoption considering constraints imposed by the ongoing interaction context (e.g., level of crowding in a room, other agents' mental states). Finally, the cognitive model gives the robot the capability to explain the reasons that led it to adopt τ at a specific level of autonomy.

According to this cognitive model, the robot possesses the capacity to adopt delegated tasks at various autonomy levels. For instance, the robot can execute the exact task delegated by the user, providing literal help. Otherwise, it can modify the user's plan, considering additional user interests or active goals. This results in the adoption of a more complex form of assistance, such as critical help or over help. Sub-help, on the other hand, occurs every time the robot fails to accomplish the task adopted in literal help, but still manages to achieve a sub-goal related to the task.

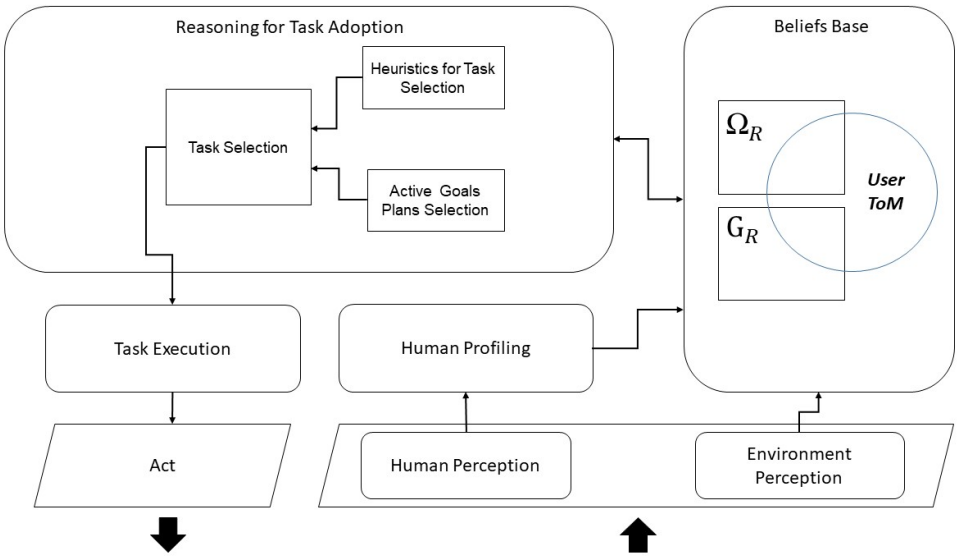

Figure 1. Perception–reasoning–action (PRA) cycle of the cognitive architecture. Please note that every module can manipulate (read or write) G_R and Ω_R.

3. Related Works

In this section, we conduct a comprehensive examination of the literature on two distinct levels. The first level focuses on adaptivity within HRI, which serves as the benchmark against which we compare our proposed theoretical model, adjustable social autonomy. The second level of analysis pertains specifically to the application domain explored in our experiment; namely, HRI within museum settings.

3.1. Adaptivity in HRI

One of the most significant challenges in the field of social robotics, and more broadly, in HRI, lies in the development of adaptive autonomous robots for real-world scenarios. Researchers have proposed various remarkable applications [7,11,16] in different domains for adaptive social robots, which may not necessarily be autonomous but possess the ability to perceive user information (e.g., cognitive or physical profile, emotions, past interactions) and make decisions accordingly. These robots can also exhibit other adaptive capabilities, such as understanding and expressing emotions, engaging in high-level communication, learning from user feedback, establishing social relationships, and reacting to various social situations. Hoffman and collegues [17] proposed a robotic teammate able to exploit an educated anticipatory action selection, based on expectations of each other's behavior. They conducted a comparative analysis of the performance of the selection process, with the performance of a reactive agent. Results showed that participants performed significantly better in the adaptive anticipatory case compared to the reactive case. Tapus et al. [18] designed a behavior adaptation robotic system capable of adjusting social interaction parameters toward customized post-stroke rehabilitation therapy, based on the user's personality traits and task performance. Findings demonstrated how the social robot's autonomous behavior adaptation to the user's personality resulted in improved human task performance. Belpeame et al. [19] introduced a personalized child–robot interaction model based on the robot's ability to model the child. This information was used by the decision-making process for reasoning about the goals of the activity and the behavior of the child. Evaluation iterations highlighted the potential of the methodology in terms of the robot's adaptation to children and the resulting influences on their engagement. Devin and Alami [20] presented a framework for social robots integrating ToM in robot decision-making. The system permitted the robot to adapt to humans' changing decisions and to provide only necessary information to them, without being intrusive, achieving a fluent execution of a shared plan. The architecture relied on multi-modal perception to

infer mental states and achieved successful collaborative pick and place tasks. Lemaignan et al. [21] proposed a cognitive architecture for HRI that allowed a robot to share space and tasks with humans, and to interact in order to support the human's actions and decisions. The model provided mechanisms for the robot to reason about the mental states of its human partners, in a scenario where a human and robot shared the same task to achieve. Görür et al. [22] integrated ToM into a robot's decision-making process for collaborative tasks. The robot was able to infer human intentional and emotional states during shared plan execution and adapt to changes of such internal states in order to behave appropriately. Umbrico et al. [23] introduced a cognitive architecture dedicated to supporting the intelligent and adaptive behavior of social robots. The scenario was a training service where the robot is used for delivering a cognitive rehabilitation program to different patients, characterized by different cognitive profiles, and assessed by a clinician. Tanevska et al. [24] proposed a cognitive architecture enabling a robot to adjust its behavior to suit different interaction profiles, using its internal motivation, which guided the robot to engage and disengage from interaction accordingly, while also taking into account the behavior of the person interacting with it. The work explored adaptation in the context of free-form social human–robot interaction. Vinanzi and Cangelosi [25] proposed CASPER, a cognitive architecture based on both symbolic and sub-symbolic paradigms to perform intention reading and collaboration in HRI scenarios. The architecture was tested in a simulated kitchen environment and the results show the robot's ability to both recognize an ongoing user's goal and to properly collaborate with them. Maroto-Gomez et al. [26] proposed a social robot able to produce a personalized interactive communication experience by considering the preferences of the interacting user. An HRI experiment was conducted in order to investigate whether users perceived the personalized selection of activities more appropriately than a random selection. The results confirmed that hypothesis, improving user likeability towards the robot. Irfan et al. [27] presented a personalized socially assistive robot for the outpatient phase of a long-term cardiac rehabilitation program. Personalization features, such as recognizing patients, addressing them by their name, tracking their attendance, and providing progress feedback were developed to improve and maintain motivation over the long-term interaction. As mentioned earlier in this paper, we adopted an approach that systematically applies the theory of adjustable social autonomy as formalized in [8]. To the best of our knowledge, a cognitive system explicitly based on this theoretical characterization, which includes multiple levels of autonomy for adopting a delegated task, and investigates complementary spaces of action in social interaction, represents a novelty in HRI computational and theoretical models.

3.2. Adaptive Social Robots in Museum Settings

Cultural heritage sites such as museums represent a complex scenario, suitable for robots that are able to assist and interact with people in a natural manner. Different pioneering work [28–32] has been proposed, with the goal of designing robots able to be deployed in a museum and to integrate different HRI capabilities. Despite their pioneering and remarkable work in autonomous social navigation and interaction with humans, these approaches do not realize a real behavior adaptation [33] to the user experience. Multiple works [34] proposed technical methods to integrate social navigation, user perception, and verbal and non-verbal communication in order to adopt different behaviors on the basis of the users involved in the interaction. Recent works [35–37] tried to implement a much more effective and personalized user experience by designing robots that are able to exploit their perceptive and decision-making skills to establish much deeper interactions with visitors. Our approach is to try to achieve a real user's adaptation, where the robot engages in reasoning based on the task delegated by the user. This reasoning process is grounded in complex models of both the interacting user (theory of mind) and the interaction mode, as described in the introduction of this work.

4. Methodology

To assess the impact of the adjustable social autonomy paradigm on HRI, we conducted an experiment in a real museum. The present experiment involved interactions between human users (museum visitors) and the humanoid robot Nao, acting as a museum assistant. During the experiment, Nao provided museum visitors the option to embark on a virtual museum tour through a computer monitor at the museum's exit. The user-assigned task to the robot was to provide assistance by offering a guided museum tour that showcased artworks from an explicitly preferred artistic period. Notably, the artworks featured in the virtual tour were unrelated to those in the actual museum exhibition. Nao's primary goal was to assist visitors in selecting the museum exhibition to visit. Subsequently, Nao guided users through the selected tour.

The present experiment focused on analyzing the impact of intelligent help on visitor satisfaction and the level of surprise as indicators of the robot's ability to intercept the visitors' needs, even when not explicitly declared. In particular, we investigated the visitors' satisfaction with (1) the suggested tour (adoption results) and (2) the robot's behavior (adoption strategy). Moreover, our experimental design explored two additional dimensions within the HRI field: the robot's trustworthiness and the role of explainability. However, these aspects exceed the scope of the present article and will be addressed in future work (see Section 8).

4.1. Definition of Adoption Levels

As mentioned earlier, the robot employed different levels of task adoption in order to offer optimal support to the user. Specifically, two types of help were exploited in this experiment: literal help and critical help. When the robot opted for literal help, it constructed the tour by selecting the most relevant artworks from the artistic period explicitly indicated by the user. Conversely, with critical help, the tour was crafted based on broader criteria, considering additional user profile information such as tolerance for room crowding or disinterest in specific artistic periods. The robot considered the potential user's interest in highly relevant artworks within the virtual museum, not necessarily belonging to the artistic period specified in the task delegation; this aspect was implicitly inferred by the robot and not explicitly stated by the user. Specifically, the robot aimed to optimize the correlation between the relevance of the artworks and the simulated room crowding where they were placed. The strength of this form of assistance lies precisely in the robot's ability to go beyond the goals declared by the user and address other needs and interests that the user may not immediately consider. However, it is crucial to note that this type of assistance is susceptible to potentially erroneous robot interpretations and could be met with reluctance from the user regarding a tutoring role played by the robot, which was not directly requested.

4.2. Hypothesis

We aimed to verify the following research hypotheses:

$\boldsymbol{H_1}$: user satisfaction regarding the quality of the tour suggested by the robot was higher when the robot provided critical help, as opposed to literal help.

$\boldsymbol{H_2}$: users exhibited greater satisfaction with the robot's behavior when the robot operated in critical help, rather than providing literal help.

$\boldsymbol{H_3}$: users experienced a higher level of surprise with the robot's selection when it performed critical help compared to literal help.

Although $\boldsymbol{H_1}$ specifically pertains to satisfaction regarding the robot's performance (suggested tour), $\boldsymbol{H_2}$ aims to investigate participant satisfaction with the robot's behavior. This behavior is intended as the decision-making process that led to the adoption of the task by performing either literal help or critical help. The participants precisely evaluated how

the robot operated concerning the task initially delegated to it, regardless of the satisfaction linked to the outcome obtained following its behavior.

4.3. Participants

We recruited 84 participants for the purpose of this study. In order to reduce ecological concerns, the experiment was conducted in an authentic museum setting, namely the Palazzo delle Esposizioni [38], located in Rome. The user sample consisted of 49 men and 35 women. A total of 28 of the participants ranged between 18 and 35 years old, 54 between 36 and 65 years old, and 2 were older than 65 years old. Participants reported their level of expertise in the artistic domain using a 5-item Likert scale ranging from lower expertise to maximum expertise (M = 2.655, SD = 0.857). Prior to their participation in this study, each participant approved an informed consent.

4.4. Design

We conducted a human-participant experimental study in a within-subject fashion. We implemented two experimental conditions in a counterbalanced order: in the control condition (LH), the robot provided literal help to the user. During the experimental condition (CH) the robot performed critical help (see Section 4.1 for literal help/cricital help definitions).

Figure 2 illustrates the scenario's design, which was specifically created in a dedicated room of the Palazzo delle Esposizioni. After visiting the exhibitions available in the building, each user had free access to this section in order to participate in the experiment. As the figure shows, the robot is in front of the user, near a screen displaying a web interface.

Figure 2. Experimental scenario representation. Not an actual participant.

Through this interface, the user can respond to questions about her artistic preferences posed by the robot and explore the virtual tour (Figure 3). The virtual tour consists of a selection of artworks, selected by the robot, according to the criteria exposed in Section 4.1.

In addition to guiding the user through various phases of the experiment, the robot describes (via speech) each artwork, providing different levels of detail depending on the user's initial preferences. During the visit, the user may encounter noisy conditions, similar to those experienced during a real museum visit. For instance, artworks may be corrupted by simulated noise to mimic the presence of other users viewing the same artwork (Figure 4). Additionally, the robot's descriptions of the artworks may be superimposed on background noise of varying intensity, simulating the crowded environment of the room where the user is viewing the artwork.

Figure 3. The figure shows an example of how an artwork is exposed to the user. (Image credits: Galleria Borghese/photo Luciano Romano).

Figure 4. Noise simulation in the artwork. (Image credits: Galleria Borghese/photo Luciano Romano).

The investigation of user satisfaction and their level of surprise involved the following experimental phases:

1. *Starting Interaction*: the robot started the interaction by introducing itself to the user, providing an overview of its role and the virtual museum it managed;
2. *User Profiling*: The robot investigated the user's artistic preferences in terms of favored and less favored artistic periods. Subsequently, the robot asked the user about the desired level of detail in the artwork descriptions as an indicator of the user's intended level of engagement. Lastly, the robot asked the user to specify her tolerance level regarding the number of people present in a room while viewing an artwork (crowding index).
3. *First Tour presentation*: After finalizing the user's profile, the tour composition was determined by the condition the robot randomly decided to implement (LH or CH). Once the selection was complete, the robot activated the corresponding tour and left the control to the user, allowing her the virtual tour.
4. *First Tour evaluation*: at the conclusion of the first virtual tour, the robot administrated a questionnaire to assess the user satisfaction and surprise level with the completed visit (see Section 4.6).
5. *Second Tour presentation*: subsequently, the robot offered an alternative tour, implementing the opposite type of help compared to the one randomly selected in the first tour.

6. *Second Tour evaluation*: after the completion of the second tour, the robot administrates the user with a new satisfaction/surprise questionnaire.

4.5. Materials

For the robot's decision-making system, we utilized an agent-oriented programming (AOP) software known as Jason [39,40], a widely adopted tool for programming artificial agents. The computational model underlying the robot's decision-making process is detailed in [3]. To develop the web application (i.e., virtual museum), we opted for the Java-based Spring Boot framework [41]. Furthermore, the robot employed in this study is the humanoid Nao robot [9], extensively used in HRI scenarios. The Nao robot operates on a specialized Linux-based operating system known as NAOqi, which powers the robot's multimedia capabilities. These capabilities include four microphones for voice recognition and sound localization, two speakers for multilingual text-to-speech synthesis, and two HD cameras for computer vision tasks such as facial and shape recognition. We collected a mySQL database of 344 artworks, organized in artistic periods in such a way that it covers the entire body of the history of art. The categorization of the history of art periods is based on the work of one of the most important art historians of the 20th century, Giulio Carlo Argan [42]. Artworks are evenly distributed among all artistic periods. All the employed images had a creative commons license and could be uses for any purpose.

4.6. Measurements

As outlined in Section 4.4, participants responded to a questionnaire administered at the conclusion of both control and experimental conditions. Through this questionnaire, we aimed to measure the following variables:

1. *Tour Satisfaction*: User satisfaction with the presented tour was measured using the following question: "How satisfied are you with the quality of the visit that the robot offered you?" Participants responded using a 5-item Likert scale ranging from not satisfied at all (1) to completely satisfied (5);
2. *Robot Behavior Satisfaction*: For assessing user satisfaction with the robot's behavior, participants were asked the following question: "How satisfied are you with the choice of robot compared to the artistic habits you expressed at the beginning of the interaction?" Participant satisfaction was investigated the same as with Tour Satisfaction;
3. *Surprise*: For assessing the users' surprise about the type of help provided by the robot, we exploited the question "How surprised are you by the choices made by the robot?" Participants responded using an 11-item Likert scale. A rating of -5 indicated a strong negative surprise, 0 represented no surprise at all, and 5 signified complete positive surprise.

5. Statistical Analysis

The statistical analyses were carried out with IBM SPSS 27.0 [43] and R through Jamovi [44]. A significance level of $\alpha = 0.05$ was selected for all analyses and p-values were reported. When significant effects were identified, the results include the unstandardized estimate (β), the 95% Confidence Intervals (CI), and the standard error of the estimate (SE). Before running the analyses, we assessed the normality of the model-dependent variables (i.e., Tour Satisfaction) using the Shapiro–Wilk test. The results highlighted statistical significance (all p-values < 0.001), suggesting a substantial deviation from normality. More specifically, both Tour Satisfaction and Robot Performance Satisfaction were negatively skewed (Skewness = -0.24; -0.09). For these reasons, we employed Generalized Mixed Effect Models [45]. In the case of skewed positive values, GLMM analysis was executed using the GAMLj module [46] using a Gamma distribution with a log link function [47]. These models reported the best Akaike's and Bayesian Information criteria (AIC and BIC) as compared with GLMM built using identity and inverse link functions. Lastly, all the models included participants' intercepts and slopes as random effects. This approach

accounted for the inherent variations wherein participants could exhibit diverse baseline levels of Satisfaction and could show singular changes in Satisfaction as a function of the Help Type.

6. Results

The main descriptive statistics are summarized in Table 1.

Table 1. Descriptive statistics of the main dependent variables.

	Literal Help			Critical Help		
	N	**Mean**	**SD**	**N**	**Mean**	**SD**
Tour Satisfaction	84	3.17	1.00	84	3.52	1.11
Robot Performance Satisfaction	84	3.82	1.04	84	3.48	0.94
Surprise	84	3.25	0.83	84	3.49	0.88

To assess the influence of Help Type (literal help or critical help) on users' self-reported Tour Satisfaction ($\boldsymbol{H_1}$), we conducted a Generalized Linear Mixed Model (GLMM) with a Gamma distribution and a logarithmic link function (see Figure 5). Help Type and Presentation Order (i.e., literal help first or critical help first) were modeled as fixed factors, whereas the Robot Behavior Satisfaction score was included as a covariate. Additionally, we incorporated individual variability by including participants' intercepts and slopes as random factors. In our model, Help Type exhibited a positive impact ($\beta = 0.11, SE = 0.03, 95\%CI[0.05, 0.17], p < 0.001$) on Tour Satisfaction, indicating that a specific robot strategy (critical help) positively influenced satisfaction levels. Similarly, Robot Behavior Satisfaction had a significant positive effect ($\beta = 0.16, SE = 0.02, 95\%CI[0.12, 0.21], p < 0.001$), suggesting that greater satisfaction with the robot's behavior during the visit predicted a higher Tour Satisfaction. In contrast, the main effect of Presentation Order was not significant ($p = 0.91$). However, we detected a significant interaction effect between Help Type and Presentation Order ($\beta = 0.17, SE = 0.05, 95\%CI[0.06, 0.03], p = 0.003$), indicating that the impact of Help Type on Tour Satisfaction depended on the order in which it was presented during the visit experience. Subsequent Bonferroni-corrected post hoc analyses revealed a statistically significant increase in Tour Satisfaction exclusively among participants who received critical help followed by literal help ($p < 0.001$), compared to those who received literal help followed by critical help ($p = 1.00$) (Figure 5).

Secondly, to investigate the impact of Help Type on Robot Behavior Satisfaction ($\boldsymbol{H_2}$), we conducted a GLMM using a Gamma distribution and a logarithmic link function (see Figure 6). This analysis also included Presentation Order and the Help Type × Presentation Order interaction as fixed factors. The analysis revealed a significant effect for Help Type ($\beta = -0.09, SE = 0.03, 95\%CI[-0.15, -0.04], p < 0.001$), indicating that the critical Help Type had a negative effect on Robot Behavior Satisfaction compared to the literal help type. Conversely, both the main effect of Presentation Order and the interaction effect were not statistically significant ($p = 0.149$; $p = 0.242$).

Finally, to examine the surprise levels after receiving different types of help ($\boldsymbol{H_3}$), we performed a Generalized Linear Mixed Model (GLMM) (see Figure 7). In this analysis, Help Type and Presentation Order were treated as fixed factors, and Robot's Performance Satisfaction was included as a covariate. As in the previous models, participants' intercepts and slopes were considered random factors. Firstly, Robot Performance Satisfaction, representing the perceived satisfaction for the robot's suggestions during the visit, exhibited a significant positive effect ($\beta = 0.88, SE = 0.20, 95\%CI[0.49, 1.26], p < 0.001$) on the participants' surprise. This finding suggests that greater satisfaction with the robot's performance was associated with higher levels of surprise. Additionally, Help Type reported a significant positive effect ($\beta = 0.76, SE = 0.29, 95\%CI[0.21, 1.33], p = 0.009$), proving that

participants who received critical help reported experiencing more surprise compared to those who received literal help. Furthermore, Presentation Order displayed a significant negative effect ($\beta = -0.92, SE = 0.36, 95\% CI[-1.63, -0.21], p = 0.013$). This suggests that the sequence in which the interaction occurred (between critical and literal help) influenced surprise levels, with lower levels of surprise observed when critical help preceded the literal one. Lastly, the interaction effect between Help Type and Presentation Order did not reach statistical significance ($p = 0.140$). In summary, this analysis unveiled distinct effects on participants' levels of surprise. Specifically, higher levels of Robot's Performance Satisfaction and receiving critical help were associated with increased levels of surprise, whereas receiving critical help at the beginning of the experiment reduced surprise scores.

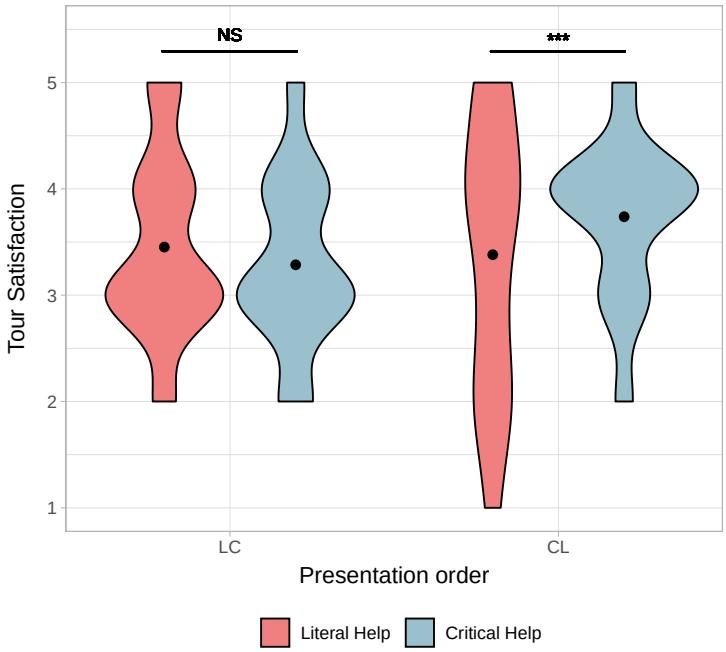

Figure 5. Tour Satisfaction as a function of Help Type and Presentation Order. *** $p < 0.001$; NS, Not Significant.

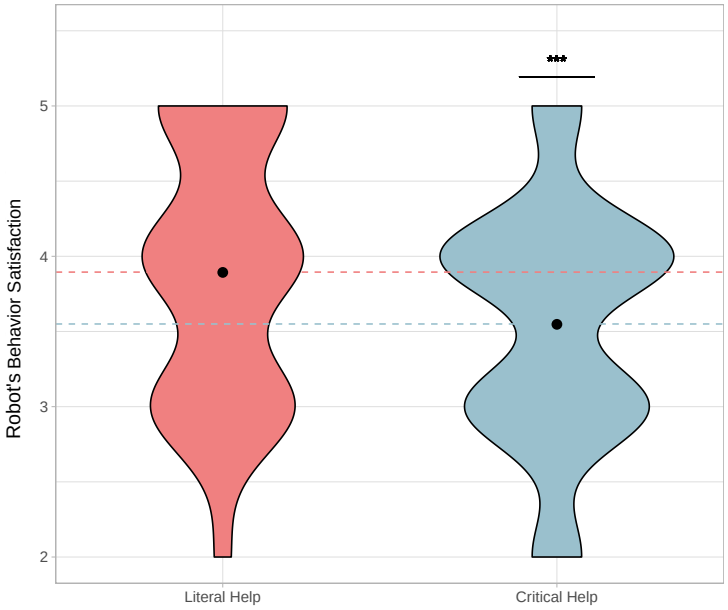

Figure 6. Robot Behavior Satisfaction as a function of Help Type. The violin plot visualizes the distribution curve. The horizontal dashed lines represent the median values. *** $p < 0.001$.

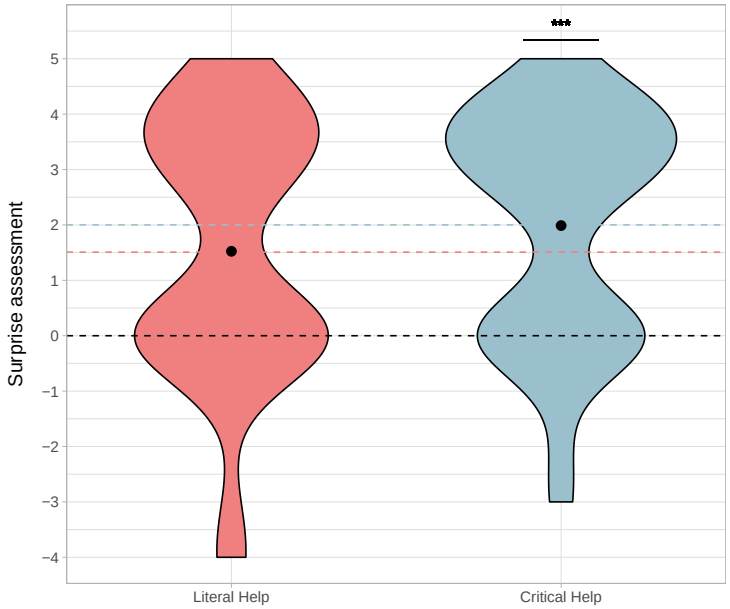

Figure 7. Surprise as a function of Help Type. The violin plot visualizes the distribution curve. The horizontal dashed lines represent the median values. *** $p < 0.001$.

7. Discussion

As exposed in the hypotheses defined in Section 4.2, our main research interest was to understand the effect of a robot employing its critical autonomous capacity in interactions with humans. In this paper, our investigation mainly focused on two key dimensions:

1. The effect on user satisfaction: we aimed to discern how the introduction of critical capacity affects user satisfaction ($\boldsymbol{H_1}$ and $\boldsymbol{H_2}$);
2. The level of surprise: we sought to determine the extent of surprise elicited in humans by this type of robot behavior ($\boldsymbol{H_3}$).

Our findings confirmed $\boldsymbol{H_1}$: user satisfaction with the quality of the proposed tour was greater when the robot developed a non-trivially literal solution, thereby introducing what we have defined as critical help, compared to a delegation aligned with the user's artistic preferences; namely, literal help. This finding was independent of the order in which the type of help was provided.

Hypothesis $\boldsymbol{H_2}$ focused on user satisfaction with the robot's performance in two types of task adoption: literal and critical help. Contrary to our expectations, the results revealed a preference towards the robot when it provided literal help rather than critical help, regardless of the order in which these levels of task adoption were introduced. It is essential to consider the following factors: (i) the average scores of user satisfaction for both literal and critical help were relatively high, even though the literal help received a higher rating; (ii) these results were not affected by the explanation that the robot provides after the interaction. The potential impact of these explanations will be further examined in a future study.

Regarding hypothesis $\boldsymbol{H_3}$, which concerned the level of user surprise in response to the robot's adoption strategies, results showed that users exhibited a relatively high level of surprise in both instances of literal and critical help. However, hypothesis $\boldsymbol{H_3}$ was confirmed as users showed greater surprise when they received critical help from the robot. The high degree of surprise even in the case of literal help likely stems from the intrinsic nature of the Human–Robot Interaction and the unexpectedly positive outcome arising from the robot's ability to select artworks based on the user's artistic preferences. This could also be explained by the fact that higher levels of satisfaction with the robot's behavior were associated with increased surprise.

Our results, taken as a whole, represent initial experimental evidence of the impact of the adjustable social autonomy model in HRI scenarios.

The robot's capacity to attribute mental states to the user allows it to dynamically adapt the task delegated to the user's unexpressed needs. This is suggested by the fact that user satisfaction increases when they receive a type of museum tour that not only considers the artworks belonging to their preferred period, but also includes relevant works from other similar periods that were selected based on the user noise tolerance level. This represents an initial confirmation of the importance of introducing autonomous robots in terms of cognitive agents, namely agents possessing propositional attitudes, assuming an intentional stance, and having a representation of the other agent's mind. An autonomous robot that adapts its behavior may not be able to tailor it to a user's needs but may instead adopt other criteria, such as considering only the resources available in the physical world. This could lead the robot to make choices that potentially may have nothing to do with the user's expectations. This would result in user dissatisfaction with the outcome achieved by the robot. What we observed in our experiment is that a robot capable of adapting its behavior to the user's needs, including implicit mental states, can adopt a task in a way that does not decrease user satisfaction compared to the result obtained.

Assistance that goes beyond literal interpretation satisfies the user more than a kind of help that merely considers what the user has explicitly expressed. These data represent further confirmation of the adjustable social autonomy model: robots that adopt tasks at different levels of autonomy change their role from being systems capable of executing even complex tasks to becoming collaborators (i.e., agents who take the initiative to provide unexpected behavior that diverge from the delegated task, either by proposing or directly executing different actions.

A higher satisfaction with the robot's behavior when it provides literal help can be attributed to the inherent risks associated with assistance that goes beyond explicit delegation, as it can generate misunderstandings, doubts, susceptibility, and refusals. This observation is corroborated by the finding that users are much more satisfied when the robot provides literal help. This confirms the potentially conflictual nature of collaboration when the robot adjusts its level of task adoption. This might also generate the perception of uncertainty regarding the endorsement of an autonomous robot's behavior. In this sense, it is relevant for the robot to evaluate the risks of conflicts that could arise due to inappropriate initiatives. Nevertheless, the tour satisfaction is higher in critical help than in literal help. In other words, the user exhibits higher satisfaction with the final result proposed by the robot when the robot adopts autonomy criteria in its collaborative efforts. However, the user may experience some degree of perplexity when assessing a robot that chooses to collaborate with a certain level of autonomy. Another confirmation of the positive yet equally conflicting impact of critical help lies in the fact that users are positively surprised by critical help.

8. Conclusions and Future Works

In this work, we presented an approach to HRI based on adjustable social autonomy. This paper describes how adjustable social autonomy can be used to systematically characterize the behavior of a robot when it adopts a task explicitly delegated by a human. The core of this approach is the robot's capability to attribute mental states to the delegating user and execute the task at various levels of autonomy. This is achieved by considering the user's mental states (goals, beliefs, expectations, and so on) that go beyond what is explicitly delegated. We tested the impact of this approach on user satisfaction and surprise, and although there are limitations to this experiment, the results are promising, demonstrating how this type of interaction can be both highly satisfying and surprising for the user. However, it is important to note that it can also lead to collaborative conflicts that may affect user satisfaction with the robot's behavior. These collaborative conflicts can be mitigated using various strategies. Results and their discussion point out the potential of exploiting adjustable social autonomy as a paradigm for developing computational cognitive models

for adaptive social robots that can improve user experiences in multiple HRI real domains. Certainly, experiments of this nature should be conducted in various domains of social robotics to validate the robustness of this type of interaction across different application domains. This aspect will be one of the points that we will need to address in future work. Another key area of future work will focus on exploring the impact of the robot's ability to provide explanations for its adopted strategy in relation to the delegated task. This will involve the analysis of data collected within the same experiment, which was not addressed in this article. Finally, a significant future work will involve investigating the effect of the approach described in this study on the trustworthiness that humans attribute to a robot capable of integrating the principles of adjustable social autonomy. This research will encompass an analysis of the results obtained during the same experiment described in this work, as well as the implementation of new experiments aimed at revealing the relationship between trustworthiness and intelligent help. What we tried to achieve in this work was to focus on autonomy, framing the paper in relation to a robot's ability to adapt to user needs and demonstrate the potential of adjustable social autonomy. Considering trustworthiness and explainability would have required framing this work in a different state of the art than what we addressed in this study. Therefore, we decided to defer the analysis of the role of explainability and trustworthiness in intelligent help to future work. We plan to describe both the theoretical models and the literature more specifically. This strategic approach allows focused exploration of specific research themes, preventing information overload and fostering in-depth understanding. We believe this is a prudent choice given the complexity of the topics we are trying to address and the extensive literature associated with them.

Author Contributions: Conceptualization, F.C. and R.F.; data curation, F.C. and M.M.; formal analysis, M.M.; funding acquisition, R.F.; investigation, F.C.; methodology, F.C. and R.F.; project administration, R.F.; software, F.C.; supervision, R.F.; validation, F.C., R.F. and M.M.; visualization, F.C. and M.M.; writing—original draft, F.C.; writing—review and editing, F.C., R.F. and M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FAIR—Future Artificial Intelligence Research (MIUR-PNRR).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available as this study belongs to an ongoing research project.

Acknowledgments: We would like to thank the management of the Palazzo delle Esposizioni for giving us the precious opportunity to carry out this experiment in a real scenario. We would like to thank Cristiano Castelfranchi for their suggestions and the productive discussion on the research topic.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mahdi, H.; Akgun, S.A.; Saleh, S.; Dautenhahn, K. A survey on the design and evolution of social robots—Past, present and future. *Robot. Auton. Syst.* **2022**, *156*, 104193. [CrossRef]
2. David, D.; Thérouanne, P.; Milhabet, I. The acceptability of social robots: A scoping review of the recent literature. *Comput. Hum. Behav.* **2022**, *137*, 107419. [CrossRef]
3. Cantucci, F.; Falcone, R. Collaborative Autonomy: Human–Robot Interaction to the Test of Intelligent Help. *Electronics* **2022**, *11*, 19. [CrossRef]
4. Gasteiger, N.; Hellou, M.; Ahn, H.S. Deploying social robots in museum settings: A quasi-systematic review exploring purpose and acceptability. *Int. J. Adv. Robot. Syst.* **2021**, *18*, 17298814211066740. [CrossRef]
5. Ragno, L.; Borboni, A.; Vannetti, F.; Amici, C.; Cusano, N. Application of Social Robots in Healthcare: Review on Characteristics, Requirements, Technical Solutions. *Sensors* **2023**, *23*, 6820. [CrossRef] [PubMed]
6. Fang, S.; Han, X.; Chen, S. The impact of tourist–robot interaction on tourist engagement in the hospitality industry: A mixed-method study. *Cornell Hosp. Q.* **2023**, *64*, 246–266. [CrossRef]
7. Ahmad, M.I.; Mubin, O.; Orlando, J. A systematic review of adaptivity in human–robot interaction. *Multimodal Technol. Interact.* **2017**, *1*, 14. [CrossRef]

8. Falcone, R.; Castelfranchi, C. The human in the loop of a delegated agent: The theory of adjustable social autonomy. *IEEE Trans. Syst. Man-Cybern. Syst. Hum.* **2001**, *31*, 406–418. [CrossRef]
9. Robaczewski, A.; Bouchard, J.; Bouchard, K.; Gaboury, S. Socially assistive robots: The specific case of the NAO. *Int. J. Soc. Robot.* **2021**, *13*, 795–831. [CrossRef]
10. Castelfranchi, C.; Falcone, R. Towards a theory of delegation for agent-based systems. *Robot. Auton. Syst.* **1998**, *24*, 141–157. [CrossRef]
11. Bianco, F.; Ognibene, D. Functional advantages of an adaptive theory of mind for robotics: A review of current architectures. In Proceedings of the 2019 11th Computer Science and Electronic Engineering (CEEC), Colchester, UK, 18–20 September 2019; pp. 139–143.
12. Montes, N.; Luck, M.; Osman, N.; Rodrigues, O.; Sierra, C. Combining theory of mind and abductive reasoning in agent-oriented programming. *Auton. Agents Multi-Agent Syst.* **2023**, *37*, 36. [CrossRef]
13. Cantucci, F.; Falcone, R. Towards trustworthiness and transparency in social human–robot interaction. In Proceedings of the 2020 IEEE International Conference on Human–Machine Systems (ICHMS), Online, 7–9 September 2020; pp. 1–6.
14. Rao, A.S.; Georgeff, M.P. BDI agents: From theory to practice. In Proceedings of the ICMAS, San Francisco, CA, USA, 12–14 June 1995; Volume 95, pp. 312–319.
15. De Silva, L.; Meneguzzi, F.R.; Logan, B. BDI agent architectures: A survey. In Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), Yokohama, Japan, 11–17 July 2020.
16. Rossi, S.; Ferland, F.; Tapus, A. User profiling and behavioral adaptation for HRI: A survey. *Pattern Recognit. Lett.* **2017**, *99*, 3–12. [CrossRef]
17. Hoffman, G.; Breazeal, C. Effects of anticipatory action on human–robot teamwork efficiency, fluency, and perception of team. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Arlington, VA, USA, 10–12 March 2007; pp. 1–8.
18. Tapus, A.; Ţăpuş, C.; Matarić, M.J. User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intell. Serv. Robot.* **2008**, *1*, 169–183. [CrossRef]
19. Belpaeme, T.; Baxter, P.E.; Read, R.; Wood, R.; Cuayáhuitl, H.; Kiefer, B.; Racioppa, S.; Kruijff-Korbayová, I.; Athanasopoulos, G.; Enescu, V.; et al. Multimodal child-robot interaction: Building social bonds. *J. Hum.-Robot. Interact.* **2012**, *1*, 33–53. [CrossRef]
20. Devin, S.; Alami, R. An implemented theory of mind to improve human–robot shared plans execution. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand, 7–10 March 2016; pp. 319–326.
21. Lemaignan, S.; Warnier, M.; Sisbot, E.A.; Clodic, A.; Alami, R. Artificial cognition for social human–robot interaction: An implementation. *Artif. Intell.* **2017**, *247*, 45–69. [CrossRef]
22. Görür, O.C.; Rosman, B.S.; Hoffman, G.; Albayrak, S. Toward integrating theory of mind into adaptive decision-making of social robots to understand human intention. In Proceedings of the International Conference on Human-Robot Interaction, Vienna, Austria, 6 March 2017.
23. Umbrico, A.; De Benedictis, R.; Fracasso, F.; Cesta, A.; Orlandini, A.; Cortellessa, G. A mind-inspired architecture for adaptive hri. *Int. J. Soc. Robot.* **2023**, *15*, 371–391. [CrossRef] [PubMed]
24. Tanevska, A.; Rea, F.; Sandini, G.; Cañamero, L.; Sciutti, A. A socially adaptable framework for human–robot interaction. *Front. Robot. AI* **2020**, *7*, 121. [CrossRef]
25. Vinanzi, S.; Cangelosi, A. CASPER: Cognitive Architecture for Social Perception and Engagement in Robots. *arXiv* **2022**, arXiv:2209.01012.
26. Maroto-Gómez, M.; Castro-González, Á.; Castillo, J.C.; Malfaz, M.; Salichs, M.Á. An adaptive decision-making system supported on user preference predictions for human–robot interactive communication. *User Model. User-Adapt. Interact.* **2023**, *33*, 359–403. [CrossRef]
27. Irfan, B.; Céspedes, N.; Casas, J.; Senft, E.; Gutiérrez, L.F.; Rincon-Roncancio, M.; Cifuentes, C.A.; Belpaeme, T.; Múnera, M. Personalised socially assistive robot for cardiac rehabilitation: Critical reflections on long-term interactions in the real world. *User Model. User-Adapt. Interact.* **2023**, *33*, 497–544. [CrossRef]
28. Burgard, W.; Cremers, A.B.; Fox, D.; Hähnel, D.; Lakemeyer, G.; Schulz, D.; Steiner, W.; Thrun, S. Experiences with an interactive museum tour-guide robot. *Artif. Intell.* **1999**, *114*, 3–55. [CrossRef]
29. Thrun, S.; Bennewitz, M.; Burgard, W.; Cremers, A.B.; Dellaert, F.; Fox, D.; Hahnel, D.; Rosenberg, C.; Roy, N.; Schulte, J.; et al. MINERVA: A second-generation museum tour-guide robot. In Proceedings of the 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C), Detroit, MI, USA, 10–15 May 1999; Volume 3.
30. Nieuwenhuisen, M.; Behnke, S. Human-like interaction skills for the mobile communication robot robotinho. *Int. J. Soc. Robot.* **2013**, *5*, 549–561. [CrossRef]
31. Chella, A.; Liotta, M.; Macaluso, I. CiceRobot: A cognitive robot for interactive museum tours. *Ind. Robot. Int. J.* **2007**, *34*, 503–511. [CrossRef]
32. Willeke, T.; Kunz, C.; Nourbakhsh, I.R. The History of the Mobot Museum Robot Series: An Evolutionary Study. In Proceedings of the FLAIRS Conference, Key West, FL, USA, 21–23 May 2001; pp. 514–518.

33. Lee, M.K.; Forlizzi, J.; Kiesler, S.; Rybski, P.; Antanitis, J.; Savetsila, S. Personalization in HRI: A longitudinal field experiment. In Proceedings of the 2012 7th ACM/IEEE International Conference on Human–Robot Interaction (HRI), Boston, MA, USA, 5–8 March 2012; pp. 319–326.
34. Hellou, M.; Lim, J.; Gasteiger, N.; Jang, M.; Ahn, H.S. Technical Methods for Social Robots in Museum Settings: An Overview of the Literature. *Int. J. Soc. Robot.* **2022**, *14*, 1767–1786. [CrossRef]
35. Iio, T.; Satake, S.; Kanda, T.; Hayashi, K.; Ferreri, F.; Hagita, N. Human-like guide robot that proactively explains exhibits. *Int. J. Soc. Robot.* **2020**, *12*, 549–566. [CrossRef]
36. Park, J.; Kim, J.; Kim, D.Y.; Kim, J.; Kim, M.G.; Choi, J.; Lee, W. User Perception on Personalized Explanation by Science Museum Docent Robot. In Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Sapporo, Japan, 7–10 March 2022; pp. 973–975.
37. Cantucci, F.; Falcone, R. Autonomous Critical Help by a Robotic Assistant in the field of Cultural Heritage: A New Challenge for Evolving Human-Robot Interaction. *Multimodal Technol. Interact.* **2022**, *6*, 69. [CrossRef]
38. Palazzo delle Esposizioni Roma. Available online: https://www.palazzoesposizioni.it/ (accessed on 1 May 2022).
39. Bordini, R.H.; Hübner, J.F.; Wooldridge, M. *Programming Multi-Agent Systems in Agentspeak Using JASON*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
40. Jason, a BDI Agent Programming Language. Available online: https://github.com/jason-lang/jason/releases (accessed on 1 December 2020).
41. Spring Boot. Available online: https://spring.io/projects/spring-boot (accessed on 22 January 2022).
42. Argan, G.C. *Storia Dell'Arte Italiana*; Sansoni: Florence, Italy, 1968.
43. George, D. *SPSS for Windows Step by Step: A Simple Study Guide and Reference, 17.0 Update, 10/e*; Pearson Education India: Delhi, India, 2011.
44. Jamovi Project. Available online: https://www.jamovi.org/ (accessed on 3 January 2020).
45. Stroup, W.W. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*; CRC Press, Taylor & Francis Group: Boca Raton, FL, USA, 2013.
46. Gallucci, M. GAMLj: General Analyses for the Linear Model in Jamovi (2.6.6). 2022. Available online: https://gamlj.github.io/ (accessed on 1 December 2020).
47. Ng, V.K.Y.; Cribbie, R.A. Using the Gamma Generalized Linear Model for Modeling Continuous, Skewed and Heteroscedastic Outcomes in Psychology. *Curr. Psychol.* **2017**, *36*, 225–235. [CrossRef]