What You Reward Is What You Learn: Comparing Rewards for Online Speech Policy Optimization in Public HRI

Sichao Song*
Yuki Okafuji*
CyberAgent
Tokyo Japan
song_sichao@cyberagent.co.jp
okafuji_yuki_xd@cyberagent.co.jp

Kaito Ariu
CyberAgent
Tokyo Japan
kaito_ariu@cyberagent.co.jp

Amy Koike
University of Wisconsin-Madison
Wisconsin United States
ekoike@wisc.edu

Abstract

Designing policies that are both efficient and acceptable for conversational service robots in open and diverse environments is non-trivial. Unlike fixed, hand-tuned parameters, online learning can adapt to non-stationary conditions. In this paper, we study how to adapt a social robot's speech policy in the wild. During a 12-day in-situ deployment with over 1,400 public encounters, we cast online policy optimization as a multi-armed bandit problem and use Thompson sampling to select among six actions defined by speech rate (slow/normal/fast) and verbosity (concise/detailed). We compare three complementary binary rewards–R_u (user rating), R_c (conversation closure), and R_t ($\geq$2 turns)–and show that each induces distinct arm distributions and interaction behaviors. We complement the online results with offline evaluations that analyze contextual factors (e.g., crowd level, group size) using video-annotated data. Taken together, we distill ready-to-use design lessons for deploying online optimization of speech policies in real public HRI settings.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Field studies**; *User models*.

Keywords

Human-Robot Interaction (HRI), Conversational Robots, Thompson sampling, Multi-armed Bandits (MAB), Speech Parameter Optimization, Field Experiment, Service Robots

ACM Reference Format:
Sichao Song, Yuki Okafuji, Kaito Ariu, and Amy Koike. 2026. What You Reward Is What You Learn: Comparing Rewards for Online Speech Policy Optimization in Public HRI. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3757279.3785589

*Both authors contributed equally to this research.

HRI '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3785589

1 Introduction

Online learning has increasingly been adopted in Human–Robot Interaction (HRI) to enable personalization and policy adaptation during interaction. For social robots, this extends beyond fixed, hand-tuned scripts: the robot can learn and adapt its action selection in real-time. Such on-the-spot adaptation yields concrete benefits, especially in-the-wild settings: higher task success, smoother turn-taking, better user satisfaction, and less operator burden from constant retuning. In practice, online learning in HRI is realized with multi-armed bandits (MAB) or reinforcement learning (RL), which have repeatedly proven effective for adapting robot behaviors to user preferences across preference learning, persuasive recommendation, tutoring, and socially interactive control [4, 11, 24–26, 29].

Conceptually, MAB differ from full RL in that they maximize cumulative reward over limited trials by balancing exploration and exploitation without an explicit state model. This property—together with extensions for drifting or time-varying preferences—makes MAB attractive for in-the-wild learning where users are heterogeneous and interaction conditions shift over time [8, 15, 22].

Nevertheless, most of the empirical evaluations of MAB applications to social robots remain in closed or semi-controlled settings. Representative studies include dueling bandits for exercise preference adaptation [29], an assistive companion that adapts linguistic style from explicit feedback [26], contextual bandits for personalizing educational chatbots [4], the Assistive MAB framework formalizing human–robot assistance [5], and fairness-constrained contextual bandits for multi-user allocation [8]. Even recent human-in-the-loop contextual bandits for robot-assisted feeding, while involving real users, are short-horizon and controlled rather than sustained public deployments [2].

By contrast, truly in-the-wild investigations–naturalistic, sustained deployments with walk-up users–remain scarce. An event-scale persuasive drink adviser demonstrated that bandit-driven policy selection can operate amid rapidly changing contexts [25]. More broadly, the online experimentation literature on "bandits in the wild" documents practical pitfalls (non-stationarity, delayed rewards, interference) and field-tested strategies that generalize to interactive systems [22]. These signals suggest feasibility while underscoring design debt in reward shaping, exposure control, and contextualization for public HRI.

Taken together, prior work indicates that MAB-based personalization for social robots succeeds in controlled contexts, whereas in-the-wild learning must contend with distributional shift, diverse user behaviors, and imperfect compliance with scripted flows. A central open question is therefore how reward operationalization

shapes learned behavior under these conditions: Which observable signals should be rewarded (e.g., task success, engagement, user-reported satisfaction)? How should they be combined? How should exploration be modulated without violating social or fairness constraints [8, 15, 21, 22]?

Although some studies focused on reward design compare subjective (explicit) and objective (implicit) feedback and show benefits from combining them [21], prior HRI work has largely explored these ideas in controlled settings. Examples include preference learning with explicit user ratings over repeated encounters [3]; object-fetching that asks users clarifying questions only when needed while leveraging implicit cues [34]; interactive reinforcement learning that integrates task performance (explicit) with task engagement (implicit) to drive real-time personalization [32]; socially-aware reinforcement learning that adapts a robot's linguistic style from user feedback [27]; and social navigation that jointly plans with both implicit motion-based and explicit multimodal communication [7]. While these studies demonstrate clear benefits from blending feedback modalities, demonstrations remain predominantly closed-environment, and how different reward definitions steer online learning during long-term, in-the-wild encounters remains underexplored.

This paper targets an in-the-wild commercial facility and systematically compares multiple reward designs within a Thompson-sampling (TS) bandit for the conversational speech policy of a social robot. It is known that users in commercial facilities have diverse attributes, such as age, and that their behavior during interactions varies depending on their psychological state [18], such as motivation. This makes it difficult to create clear rules for selecting actions. TS directly represents uncertainty in reward estimates and uses it to decide when to explore versus exploit. This makes it well-suited to commercial facilities where user attributes are diverse, while keeping the policy free from brittle, hand-engineered selection rules.

We deployed a service robot in a shopping mall for 12 days (over 1,400 public encounters), adapting two dimensions–*speech rate* (slow/normal/fast) and *verbosity* (concise/detailed)–while optimizing three binary rewards that capture complementary constructs: R_u (user rating), R_c (conversation closure), and R_t ($\geq$2 turns). Each interaction was logged and video-annotated for social context (e.g., crowd level and group size), enabling post-analysis of how context moderates learned policies. We report (i) TS learning and posterior arm preferences under each reward and (ii) generalized linear models quantifying arm×context interactions. In Thompson sampling, an arm is an action type. In this experiment, each arm refers to a setting of the robot's speech rate × verbosity. Based on these findings, we distill design lessons for contextual online optimization in public HRI.

We pursue two objectives in an in-the-wild, public deployment of a conversational service robot:

- **RO1:** Evaluate how alternative reward definitions steer online learning of the speech policy of a social robot. Concretely, we compare R_u (user rating), R_c (conversation closure), and R_t ($\geq$2 turns) within a TS over speech rate (slow/normal/fast) and verbosity (concise/detailed).
- **RO2:** Quantify how social context moderates outcomes and arm effectiveness using video annotations and GLMs; then translate these regularities into actionable future guidance for context-aware online optimization.

This paper contributes two complementary advances.

- We deploy a social robot in a shopping mall over a 12-day period, demonstrating how TS adapts its speech policy in real time across over 1,400 public encounters. The learned arm preferences diverge by reward design, demonstrating construct-sensitive policy selection in the wild.
- We provide video-annotated analysis that quantifies context moderation with GLMs, revealing what factors influence performance outcomes. According to the results, we distill design lessons for context-aware online optimization in public HRI.

2 Methodology

2.1 Thompson Sampling (TS)

Thompson Sampling (TS) is a Bayesian approach to the exploration/exploitation tradeoff, dating back to Thompson's 1933 proposal of selecting actions in proportion to the posterior probability that they are optimal [31]. In the past decade, TS has received strong theoretical support, including finite time regret analyses that match or approach the best achievable rates [13], and it has been widely adopted in online recommendation and advertising [6]. Comprehensive tutorials further situate TS within the bandit literature and show its extensibility to contextual, non-stationary, and constrained settings [28]. Key advantages include (i) conceptual and implementation simplicity, (ii) exploration driven by posterior uncertainty without explicit optimism bonuses, and (iii) straightforward accommodation of real world constraints and deployment realities such as delayed rewards and batched updates.

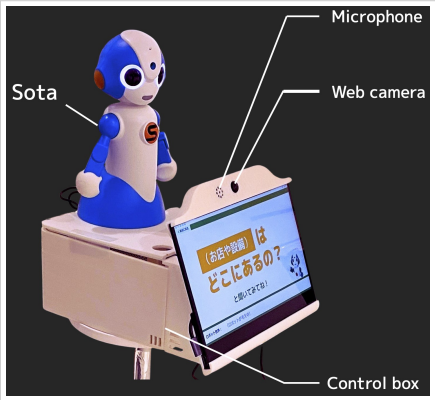
Bernoulli rewards with Beta priors. In our setting each interaction yields a binary outcome $r \in \{0, 1\}$ (success/failure), so we model the success probability of arm a as $\theta_a \sim \text{Beta}(\alpha_a, \beta_a)$. Because Beta is conjugate to the Bernoulli likelihood, observing reward r after playing arm a updates the posterior via

$$\alpha_a \leftarrow \alpha_a + r, \qquad \beta_a \leftarrow \beta_a + (1 - r).$$

At round t, TS draws a sample $\tilde{\theta}_a \sim \text{Beta}(\alpha_a, \beta_a)$ for each arm independently and selects $a_t = \arg\max_a \tilde{\theta}_a$. Arms with greater posterior uncertainty are stochastically favored, yielding principled exploration. Common uninformative initializations include α_a=β_a=1 (uniform) or a small symmetric prior. In our deployment we preceded learning with a short uniform pre-allocation to ensure minimum exposure (see §2.2). Algorithm 1 shows the standard loop for Bernoulli rewards.

We used TS to optimize among six arms (slow/normal/fast × concise/detailed). The following engineering choices aligned TS with our deployment setting:

- **Cold start.** We guaranteed minimum exposure by uniformly allocating each arm a small number of initial interactions and seeding (α_a, β_a) with those observations.



(a) Service robot system "Sota".

(b) Field setting in the shopping mall.

Figure 1: Sota robot designed for route guidance and field setting in the shopping mall. It features an integrated display. The display explicitly communicates the robot's ability to provide directional assistance, ensuring users are aware of its functionality.

Algorithm 1 Thompson Sampling for Bernoulli Rewards

Require: Arms $\mathcal{A}$; Beta priors $\{(\alpha_a, \beta_a)\}_{a\in\mathcal{A}}$

for $t = 1, 2, \ldots$ **do**
  for all $a \in \mathcal{A}$ **do**
    Sample $\tilde{\theta}_a \sim \text{Beta}(\alpha_a, \beta_a)$
  $a_t \leftarrow \arg\max_{a\in\mathcal{A}} \tilde{\theta}_a$
  Play a_t and observe $r_t \in \{0, 1\}$ (possibly delayed)
  Update: $(\alpha_{a_t}, \beta_{a_t}) \leftarrow (\alpha_{a_t} + r_t,\ \beta_{a_t} + 1 - r_t)$

- **Optional non-stationarity handling.** To discount stale evidence, a simple forgetting factor $\lambda \in (0, 1]$ can be applied to sufficient statistics: $\alpha_a \leftarrow \lambda\alpha_a + r$, $\beta_a \leftarrow \lambda\beta_a + (1 - r)$ (we report results with λ=1 as the default TS).
- **Multiple labels.** We updated the algorithm using three types of binary rewards: a third-party judgement (R_c, R_t) based on good/bad judgment, and a user rating on the site (R_u). These rewards were delivered immediately after the interaction, and updates were applied upon arrival.

Together, these practices keep TS simple while addressing non-stationarity, delayed feedback in real-world deployments.

2.2 Experiment Design

This study was conducted in collaboration with a shopping mall during normal business operations. The deployment and data handling followed an institutional ethical review and a facility agreement. The experiment was approved by the Research Ethics Committee of The University of Osaka (Reference Number: R1–5–9). This study was conducted on an opt-out basis for unwilling participants who wanted to be removed from the video data.

Field and Tasks. We conducted the field study at a shopping mall in Japan (details anonymized for double-blind review). The social robot was installed next to the mall's floor map on the second and third floors, as shown in Figure 1. Two conditions were simultaneously operated on the second and third floors, and once one condition was completed, the other condition was implemented. The robot was deployed for 12 days, approximately eight hours per day, until all experimental conditions were achieved. The robot's main task was route guidance, but it was also designed to be able to handle various types of interaction, such as casual conversation.

Reward Design. Previous research has compared subjective (explicit) and objective (implicit) feedback from users in reward design [21]. However, in real-world environments, user behavior is unique, as social robots are often used as chat partners, despite their intended role as guides. This makes it difficult to define clear objective indicators. Therefore, we firstly prepared four objective indicators to be used in social robot dialogue tasks (conversation closure, overlapping utterances, dialogue conflict, and number of turns) [12] and conducted prior observations to determine whether they were important aspects for online learning of speech policies. After preliminary observation, we decided to conduct the experiment using three types of binary rewards (one subjective and two objective) that reflected different operational goals:

- R_u **(User rating):** success if a post-interaction, single-item satisfaction rating on a 7-point scale was ≥ 6; failure otherwise.
- R_c **(Conversation closure):** success if the interaction reached an explicit conversational closing routine (farewell and/or thanks); failure otherwise.
- R_t **(≥ 2 turns):** success if the dialog contained at least two turns (user-robot exchanges); failure otherwise. An encounter statement is counted from the robot's proximity-triggered greeting, and success requires the minimal sequence R1 (robot) → U1 → R2 → U2; a "turn" is a contiguous speaker segment.

R_u captures users' subjective assessments directly, whereas R_c and R_t capture objective aspects of the interaction. For R_u, we mapped scores ≥ 6 to success. According to our previous field studies, scores were positively biased; treating 5 as "positive" would result in most interactions being classified as successful. Using the stricter ≥ 6 threshold yields a more balanced split of positive vs. negative outcomes and provides better discrimination. The remaining two objective rewards were determined by binary criteria in terms of a balanced split of outcomes: whether the conversational closing

routine happens and whether the dialogue contained at least two turns.

Our three reward conditions (R_u, R_c, R_t) operationalize this explicit–implicit feedback in a public–space deployment, enabling a controlled comparison of how each definition guides the same online learning procedure. In contrast to prior HRI formulations that relied primarily on explicit ratings [3] or combined explicit and implicit signals in lab–style studies [7, 27, 32, 34], our study evaluates these alternatives side–by–side under identical hardware, action space, and environment.

Data Collection. For each reward condition, we used a two-phase schedule: a 30-interaction cold-start phase to seed the posteriors, followed by 450 user interactions with the bandit active. Thus, we obtained 480 interactions per reward in total.

For the cold-start phase, we ran a short uniform pre-allocation: each of the six arms was pulled five times, collecting rewards and updating the per-arm posteriors. We then used the resulting Beta parameters as the bandit's initial state for Thompson sampling. This ensured a uniform minimum exposure across arms and reduced susceptibility to early mislearning.

For each interaction we logged: timestamps; the selected arm (speech rate, verbosity); rating of the arm; and bandit variables (e.g., prior/posterior parameters). These logs enable us to reconstruct learning curves and posterior summaries. In addition to logs, we recorded video footage throughout the deployment for annotation purposes.

2.3 Service Robot System

We used a small humanoid robot, "Sota" (Vstone Co., Ltd.), which is 28 cm tall and has a childlike appearance. Each hand has two degrees of freedom (DoF), enabling simple gestures (e.g., a pointing gesture to indicate "that direction"). Sota features three facial LEDs (eyes and mouth) for expressive cues and can rotate its body to realize gaze behaviors. For speech, we used Google Cloud Speech-to-Text (STT) and Text-to-Speech (TTS) APIs. Dialog content generation and dialogue-state management were supported by the OpenAI API (GPT-4 family). A front-mounted display presented short prompts and guidance, including facility maps.

As shown in Fig. 1a, a control box beneath Sota houses a mini PC that runs the behavior controller. We implemented a fully autonomous conversational system comprising four basic components: a Recognizer, Dialog Manager, Action Manager, and Modality Manager. Specifically, the Recognizer uses a 180° fisheye camera and PoseNet [14] to identify the nearest visitor and infer visit states; the Dialog Manager is a finite-state controller that greets on detection and accepts speech when nearby while grounding LLM-generated content in a curated facility knowledge base and a fixed "Sota" persona; the Action Manager composes task responses (e.g., route guidance) from dialogue history and the knowledge base; and the Modality Manager triggers about 40 word–gesture mappings—including pointing—to reinforce verbal instructions.

Thompson Sampling Integration. To enable online optimization of speech policies, we implemented a TS module that selects the arm (speech rate × verbosity) at the start of each interaction using Thompson sampling for Bernoulli rewards [6]. We designed a system in which the robot's behavior is updated in real time through interactions and evaluations. The learning flow using TS is shown in Figure 2. The reward signal was derived from two complementary sources: (i) on-site user self-reports for R_u and (ii) third-party success/failure judgments for R_c/R_t.

In case (i), users were prompted to complete a survey immediately after the interaction, and their satisfaction with the interaction was rated on a 7-point Likert scale using a survey tablet. A response of 6 or higher was a success, and otherwise was a failure, and the result was immediately sent to the robot system. In case (ii), a monitoring UI streamed camera footage from the robot. The first, second, and fourth authors independently observed interactions in real time and, after the interaction ended, pressed "Success" or "Failure" according to predefined evaluation criteria. A majority vote instantly sent success and failure labels to the robot system.

The reward results sent to the robot system immediately trigger the TS, which updates the posterior distribution of the arm used in that interaction. The robot's speech policy is then determined based on the updated posterior distribution and immediately reflected in the robot system. The updated behavior is executed the next time the user interacts. This online approach balanced exploration and exploitation, enabling real-time, on-device adaptation of speech policies in the wild.

Regarding detailed speech policy settings, for speech rate that we used Google Cloud TTS rate multipliers slow=0.80, normal=1.20, and fast=1.60; for verbosity, we injected a per-turn instruction into the GPT dialog prompt: concise = "Respond in short, concise sentences. Focus on the key points and avoid unnecessary elaboration." while detailed = "Respond in longer, more detailed prose. Include concrete examples and supplementary explanations." Before deployment, the authors piloted all six patterns (arms) on-device and reached consensus that these settings were natural, safe, and sufficiently separable for our field experiment.

2.4 Evaluation and Analysis

Bandit Performance. For each reward condition ($R_u/R_c/R_t$), we report overall success rates and posterior Beta distributions per arm to reveal learned arm preferences.

Typically, when evaluating a bandit algorithm, the cumulative reward is compared with that of uniform sampling, which performs all actions randomly as a baseline. In our experiment, in addition to running the robot under the three reward conditions, we also ran an experiment under the uniform sampling condition. However, while most of the data for the experiments under the three reward conditions was collected on weekdays, data for the uniform sampling condition was collected mainly on holidays. As a result, we concluded that a fair comparison was difficult due to significant differences in user attributes and behavior patterns. Therefore, in this study, we only compared the speech policies learned under each reward condition as mentioned in ROs, and the comparison with uniform sampling is reported in the Appendix.

Video Annotation. Apart from analyzing the behaviors acquired in response to different rewards, we also perform a post-hoc analysis to examine how various contextual factors adjust the learned speech policy, for the purpose of discussing future guidance. All 1,400+ interactions were video-recorded and coded to obtain

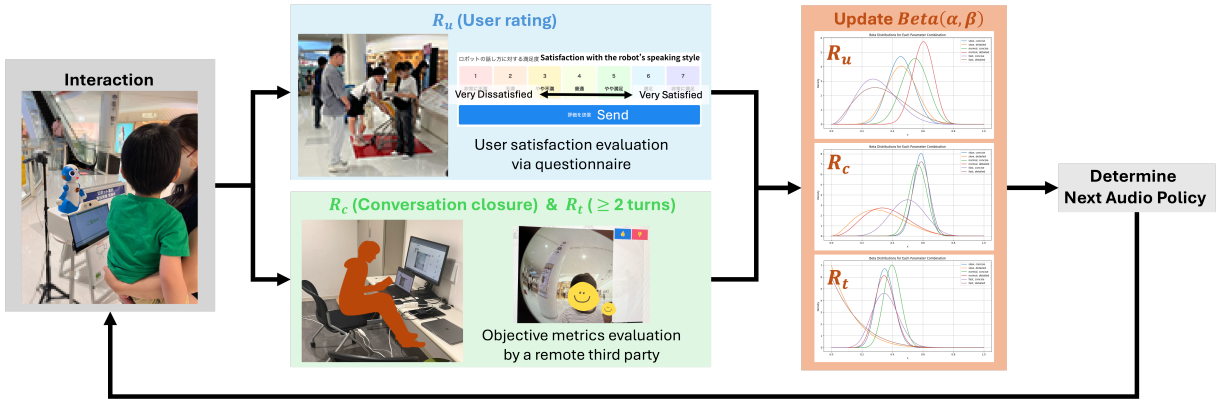


Figure 2: This diagram shows a robot's speech policy change using Thompson sampling (TS). After the interaction ends, three types of reward conditions are given. In the R_u (User rating) condition, the user who interacted with the robot evaluates the robot using a questionnaire, while in the R_c (Conversation closure) and R_t ($\geq$ 2 turns) conditions, a remote third party evaluates the objective metrics of interaction. The obtained evaluation as a reward is sent to the robot system in real time, and the robot's next behavior is determined using TS and immediately reflected.

contextual variables. We include the following factors as contextual regressors:

- **Group** (solo/group): whether the focal user approached alone or with companions. Human group presence and characteristics systematically modulate engagement and evaluations toward robots in the wild [33].
- **Crowd** (present/absent): whether bystanders were visibly near the robot during the interaction. The presence of bystanders alters helping/interaction intentions (bystander effect) [20].
- **Ask-direction** (yes/no): whether the user asked for route direction information. Wayfinding/dialogue needs (e.g., asking for directions) strongly condition dialogue strategies and outcomes in public-space HRI [10].
- **Motivation** (function/experiment/curiosity/education): the user's apparent reason for engaging with the robot. The differences in user motivation lead to differences in behavior [18].

In our open-world, multi-party setting, we deliberately exclude demographic covariates (age, gender) from the GLM. Identifying a stable main speaker is unreliable when roles and addressees shift within groups, making per-person demographics methodologically fragile. We therefore model contextual moderators that are observable and theory-motivated–Group, Crowd, Ask-direction, Motivation–rather than individual demographics.

The coding process was primarily carried out by four people, who reviewed and decided on the coding criteria in advance. Because the Group/Crowd/Ask direction are objective indicators of the situation and interaction, there was little variability in the coding criteria between coders. However, because Motivation is internal user information, the coding criteria are more likely to vary depending on the coder. Therefore, for Motivation, around 10% of the data was checked by multiple coders, resulting in a Cohen's κ coefficient of 0.53.

For the binary outcome of reward, generalized linear models (GLMs) were fitted for each reward condition using all of these interaction-coded independent variables:

$$\begin{aligned}\text{Outcome} \sim\ & \texttt{arm} + \texttt{crowd} + \texttt{group} + \texttt{ask\_direction} + \\ & + \texttt{motivation} + \texttt{arm} \times \texttt{crowd} + \texttt{arm} \times \texttt{group} \\ & + \texttt{arm} \times \texttt{ask\_direction} + \texttt{arm} \times \texttt{motivation}.\end{aligned}$$

Using these analyses, we discuss how the interaction context and user behavior may have influenced the learning outcomes of the speech policy obtained through TS. The results of the GLMs should be important for the future development of in-the-wild online learning frameworks such as MAB.

3 Results

3.1 Bandit Performance

Arm abbreviations (speech rate × verbosity): SC = Slow-Concise, SD = Slow-Detailed, NC = Normal-Concise, ND = Normal-Detailed, FC = Fast-Concise, FD = Fast-Detailed.

Overview. Table 1 reports the descriptive variables (count, success rate, and choice rate) for each reward condition. Cold starts were performed five times in each arm under each condition, except for the R_t condition, where missing data meant that not all arms were performed five times. As can be seen from the All Data values in Table 1, the arms with high success rates had high chosen rates, while the arms with low success rates had low chosen rates, demonstrating the exploration-exploitation characteristic of TS. Thus, although the R_t condition experienced missing cold starts data, it is unlikely to have had a significant impact on the overall data.

Figure 3 shows the posterior beta probabilities, summarizing the arm posterior probabilities learned across half and all data. Showing half the data allows us to see the learning process. A larger expected value of the distribution indicates a higher success rate in each reward condition. The greater the number of selections,

Table 1: Descriptive statistics by each reward (R_u, R_c, R_t) for the six audio-policy arms [SC = Slow–Concise, SD = Slow–Detailed, NC = Normal–Concise, ND = Normal–Detailed, FC = Fast–Concise, FD = Fast–Detailed]. For each reward definition, both the cold-start sample and the full-deployment sample: trials n / successes / success rate, along with the proportion selected by TS (chosen rate).

(a) R_u (user rating)

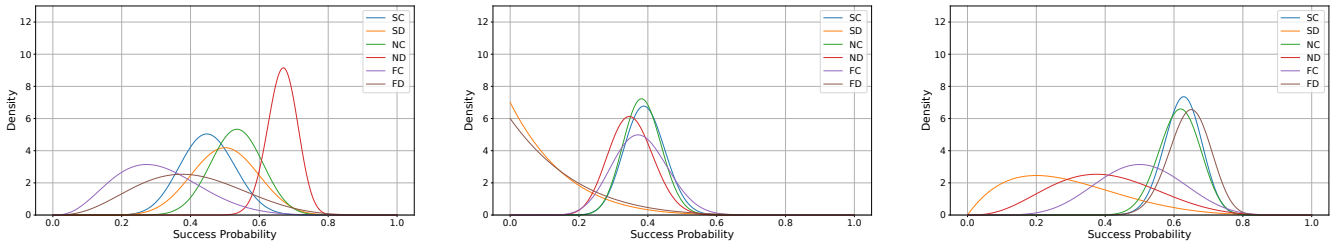
	Cold start		All data			
Arm	n	succ.	n	succ.	succ. rate %	chos. rate %
SC	5	3	44	18	40.9	9.2
SD	5	4	39	19	48.7	8.1
NC	5	3	98	62	63.3	20.4
ND	5	4	240	146	60.8	50.0
FC	5	3	16	6	37.5	3.3
FD	5	1	43	24	55.8	9.0

(b) R_c (conversation closure)

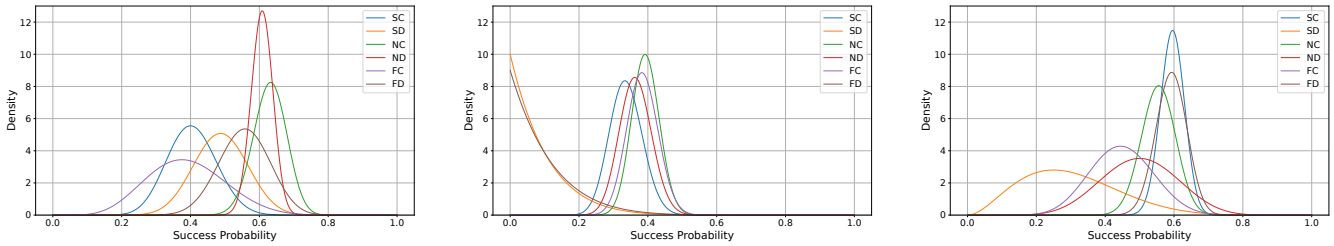
	Cold start		All data			
Arm	n	succ.	n	succ.	succ. rate %	chos. rate %
SC	5	3	96	32	33.3	20.0
SD	5	0	9	0	0.0	1.9
NC	5	4	148	58	39.2	30.8
ND	5	3	105	38	36.2	21.9
FC	5	3	114	44	38.6	23.8
FD	5	0	8	0	0.0	1.7

(c) R_t (≥2 turns)

	Cold start		All data			
Arm	n	succ.	n	succ.	succ. rate %	chos. rate %
SC	4	4	198	118	59.6	42.4
SD	3	0	8	2	25.0	1.7
NC	3	3	99	55	55.6	21.2
ND	3	1	18	9	50.0	3.9
FC	3	2	27	12	44.4	5.8
FD	3	3	117	70	59.8	25.1



(a) R_u (user rating) with half-data results **(b) R_c (conversation closure) with half-data results** **(c) R_t (≥2 turns) with half-data results**



(d) R_u (user rating) with full-data results **(e) R_c (conversation closure) with full-data results** **(f) R_t (≥2 turns) with full-data results**

Figure 3: Posterior Beta densities of per-arm success probability by each reward condition, using half and full of the training data. [SC = Slow–Concise, SD = Slow–Detailed, NC = Normal–Concise, ND = Normal–Detailed, FC = Fast–Concise, FD = Fast–Detailed]

the smaller the variance of the distribution, reflecting the results in Table 1.

R_u (User rating). A total of 480 data are analyzed. The arm with the highest chosen rate was ND (50.0%), while the arm with the highest success rate was NC (63.3%). This can be seen from the learning process in Figure 3. In the first half of the data, ND has a higher success rate than the other arms and is therefore chosen with a high probability. However, after learning all the data, NC shows the highest success rate. Therefore, NC begins to be chosen more frequently towards the second half of the data, resulting in the highest success rate. Furthermore, among the other arms, the highest chosen and success rates are FD, SD, SC, and FC, in that order, after NC and ND. Compared to the other reward conditions, the variance across all arms is small, indicating that all arms are selected most evenly across all reward conditions.

R_c (Conversation closure). A total of 480 data are analyzed. The arm with the highest chosen and success rates is NC (chosen rate: 30.8%, success rate: 39.2%). While NC performed best, SC, ND, and FC have comparable chosen and success rates, indicating that learning occurs at a rate roughly equivalent to NC from the early stages of learning. Furthermore, for SD and FD, the success rates are 0.0% at the cold start, and although they are occasionally chosen exploratorily thereafter, their success rates remain 0.0% throughout the entire data.

R_t (≥ 2 turns). A total of 467 data sets were analyzed. The highest chosen rate is SC (42.4%), and the highest success rate is FD (59.8%). This is likely due to changes in the success rate of each arm over the learning process, similar to the R_u condition. Following these two arms, the NC arm has the highest chosen and success rates, with the remaining arms showing lower chosen and success rates.

3.2 GLMs for Modulation

For each binary outcome according to the reward conditions, we fit binomial-logit GLMs with predictors and arm-context interactions. We report selected statistical summaries for $R_u/R_c/R_t$ across arms, as shown in Table 2; full GLM outputs are provided in the Appendix.

Table 2: Selected GLM effects across reward definitions (significant predictors only). Baselines: Arm = **SC**, Crowd = **no**, Group = **group**, Ask-direction = **no**, Motivation = **function**. Coefficients are GLM estimates (log-odds).

	R_u (user rating)				
Predictor	β	SE	z	p	OR
grp_single_T	-1.54	0.77	-1.99	0.05*	0.21
ND:grp_single_T	1.95	0.90	2.16	0.03*	7.03
NC:motivation_education	2.76	1.41	1.96	0.05*	15.76
ND:motivation_education	2.65	1.25	2.12	0.03*	14.18
FD:motivation_education	2.91	1.45	2.00	0.05*	18.30
	R_c (conversation closure)				
Predictor	β	SE	z	p	OR
NC:grp_single_T	2.68	1.31	2.04	0.04*	14.59
	R_t ($\geq$ 2 turns)				
Predictor	β	SE	z	p	OR
ND:crowd_T	-1.73	0.62	-2.80	0.01*	0.18
FC:crowd_T	-1.39	0.52	-2.69	0.01*	0.25
SD:motivation_education	2.33	1.16	2.01	0.04*	10.26
FC:motivation_education	1.89	0.93	2.04	0.04*	6.63

* indicates significant difference.

Coding and baselines. Arms are treatment-coded with baseline SC (Slow-Concise). `crowd_T`= 1 indicates crowd present (baseline: no crowd); `grp_single_T`= 1 indicates single user (baseline: group); `ask_directions_T`= 1 indicates asked for directions (baseline: no); `motivation` baseline is function.

R_u (user rating). Overall fit: Null deviance = 1184.7 (df = 869) decreased to residual deviance = 1103.2 (df = 828); AIC = 1187.2. Large standard errors on some arm coefficients indicate partial separation in rare arm-context strata; we focus on stable contextual contrasts and interactions.

Main effects. Single-user interactions yielded lower ratings than group encounters (p<0.05, OR=0.21). There was no overall main effect of crowd, ask-direction, or motivation.

Interactions. In single-user scenes, ND is favored relative to the SC baseline (p<0.05, OR=7.03). For education-motivated visitors, NC, ND, and FD each raise the odds of a high rating (p<0.05; OR=15.76, 14.18, and 18.30).

R_c (conversation closure). Overall fit: Null deviance = 592.5 (df = 449) decreased to residual deviance = 524.4 (df = 415); AIC = 594.4. Several interaction terms were singular (NA), indicating sparse arm–context cells.

Main effects. No main effect has a significant difference.

Interactions. A single significant interaction NC×grp_single_T (p<0.05, OR=14.59) suggests that in one-to-one encounters, Normal–Concise increased closure odds relative to SC. No other arm context interactions were significant.

R_t ($\geq$2 turns). Overall fit: Null deviance = 1218.7 (df = 891) decreased to residual deviance = 1156.3 (df = 852); AIC = 1236.3. Two interaction terms were singular (NA).

Main effects. No main-effect covariate reached significance.

Interactions In crowd scenes (relative to SC with no crowd), ND and FC are comparatively disadvantaged: ND×crowd (p<.01, OR=0.18) and FC×crowd (p<.01, OR=0.25). For education-motivated users (vs. function-motivated), SD and FC perform better than SC: SD=education (p<.05, OR=10.26) and FC×education (p<.05, OR=6.63).

4 Discussion

4.1 Summary of Results

From our 12-day in-the-wild deployment involving over 1,400 interactions, three key takeaways emerge.

Firstly, the data are not evenly distributed across arms. As posteriors concentrate, the policy favors promising arms and samples dominated ones less; moreover, arms with higher observed success rates are selected more often, consistent with Thompson Sampling's intended behavior.

Second, arm preference is sensitive to how success is defined; the online learning converged to different speech policies depending on whether success was defined as perceived interaction satisfaction (R_u), conversation closure (R_c), or conversational persistence (R_t). The results indicate that, among well-sampled arms, the best arm is NC (Normal–Concise) and ND (Normal–Detailed) for R_u, NC leads for R_c, and SC (Slow–Concise) and FD (Fast–Detailed) for R_t.

Third, social context could shape outcomes and interact with speech policies.

- Under R_u, single-user encounters rate lower than group encounters (main effect). ND especially suits single users. In education-motivated visits, NC/ND/FD tend to be rated higher.
- Under R_c, reward by conversation closure is broadly stable across context, but single-user encounters specifically favor a normal–concise delivery.
- Under R_t, in crowded scenes ND and FC underperform. By contrast, education-motivated visits favor SD and FC over other settings.

Taken together, (i) the bandit behaved as intended (higher-success arms received more pulls), (ii) the definition of "success" steers which policy emerges as optimal, and (iii) social context moderates these effects. Practically, operators should choose reward definitions aligned with their operational goal (satisfaction vs. closure vs. persistence) and deploy context-aware policies (e.g., for satisfaction, use ND for single users and consider NC/ND/FD in education-motivated cases; to maximize closure with single users, use normal–concise; to sustain conversations, avoid ND/FC in crowds and leverage SD/FC for education-motivated visitors).

4.2 Speech Policy vs. Reward Design

Our results demonstrate that the definition of a reward signal fundamentally alters the learned behavior, leading to distinct optimal policies. This underscores that there is no single "best" speech policy, but rather a policy that is optimal for a specific operational goal. The observed policy differences are consistent with reports that varying the mix of explicit and implicit user feedback can systematically alter learned behavior [3, 32]. Our contribution is to demonstrate this effect in a public deployment while directly comparing multiple reward definitions for the very same robot and action set.

For the R_u (User Rating) reward, the bandit converged on Normal-Concise (NC) and Normal-Detailed (ND) policies. This suggests that user satisfaction in this public setting is maximized by a normative and predictable interaction style. A normal speech rate is familiar and easy to process, avoiding the potential for frustration from a slow pace or the cognitive load of a fast one. The split between concise and detailed likely reflects differing user preferences for information density, but both fall within a comfortable, non-extreme range. The GLM refines this picture in two ways. First, single-user encounters yield lower ratings overall than group encounters, but ND is particularly effective for single users, consistent with the idea that a more thorough, seemingly personalized response is appreciated when no audience is waiting. Second, education-motivated visits amplify positive evaluations for NC, ND, and FD, suggesting that information-seeking users value either a clear, compact explanation (NC), a fuller step-by-step response (ND), or even fast, information-dense delivery (FD) when they are already motivated to process content.

The R_c (Conversation Closure) reward, which is predicated on reaching a clean end to the interaction, strongly favored the Normal-Concise (NC) policy. This reward operationalizes task efficiency. The NC arm provides information directly and without extraneous detail, enabling users to accomplish their goal and conclude the interaction smoothly. 0.0% success rate for Slow-Detailed (SD) and Fast-Detailed (FD) arms under this reward is telling; detailed responses likely prolong the interaction unnecessarily or introduce conversational threads that prevent a simple closure, thus failing the reward condition. The GLM sharpens this picture: NC is especially effective in one-to-one encounters.

Finally, the R_t ($\geq$ 2 turns) reward, a proxy for engagement, yielded a more complex bimodal preference for Slow-Concise (SC) and Fast-Detailed (FD). This suggests two distinct mechanisms for fostering sustained interaction. The SC policy, with its deliberate pacing, may make the robot seem more careful or accessible, potentially prompting users to ask clarifying questions or feel less rushed, thereby extending the dialogue. Conversely, the FD policy may enhance engagement through a different channel: novelty and information density. A fast, detailed response can be more entertaining and provide more conversational hooks for a user to latch onto, sparking curiosity and follow-up questions [23]. The GLM results supplement this, showing that ND and FC are comparatively disadvantaged in crowded scenes.

4.3 Design Lessons for Context-Aware Online Optimization

In this section, we argue that reward design—rather than parameter tuning alone—primarily shapes which speech policies emerge. The same reasoning extends beyond speech rate and verbosity to a broader action space (e.g., gaze, gesture, display cues), motivating a shift from average-case optimization to context-sensitive control.

Our findings, especially the strong moderation of arm performance by social context, suggest that effective online optimization in HRI must move beyond simple MAB frameworks. While the bandit learned reasonable average policies, the greater opportunity is to adapt policies dynamically to the evolving context of each encounter. Rather than prescribing fixed rules, we outline key considerations for the next generation of context-aware learning systems.

First, the richness of contextual features is paramount. Our results identified crowd level, group size, and user intent as critical variables. However, these are just a starting point. Future systems could benefit from incorporating a much wider array of contextual signals, such as the user's emotional expression [30], or even a memory of past interactions with that individual [16]. The challenge and opportunity lie in developing robust, real-time sensing capabilities to capture these nuanced features and represent them in a way that is meaningful for a learning algorithm. This moves the design focus from merely selecting an action to deeply understanding the situation in which the action is taken.

Second, we should explore more sophisticated models for policy learning. While contextual bandits are a natural next step, the dynamic and often unpredictable nature of public HRI may call for even more advanced approaches. For instance, models that can handle non-stationarity—the fact that the best policy might change over the course of a day as mall traffic patterns shift—are essential for long-term deployments [19]. Furthermore, as robots become more capable, their actions will start to influence the subsequent state of the interaction, a condition that traditional bandits do not model. This suggests a future trajectory towards full reinforcement learning (RL), where the robot learns not just an immediate action-reward link but a long-term strategy for interaction [17].

Third, the definition and delivery of the reward signal itself present a design space for exploration. Our study compared three distinct, immediate rewards. However, in long-term interactions, success might be better defined by metrics that unfold over time, such as repeat engagement or successful task completion across multiple encounters. This requires frameworks that can handle delayed or sparse rewards [1, 19]. Moreover, incorporating human feedback more directly into the reward function, for example, by learning from preferences or corrections, could allow for more aligned and personalized robot behavior [9].

This work serves as a stepping stone, demonstrating that context is not just a moderating factor but the very foundation upon which truly adaptive and intelligent social interaction should be built. The goal is not just to find the single best policy, but to create a system that can fluidly generate the right policy at the right time.

4.4 Limitations

This study took place at a commercial mall with one robot platform and six pre-defined speech policies (speech rate×verbosity). Findings may differ with other factors, such as type of voice, acoustics, or task settings. We also chose binary rewards for simplicity and sample efficiency. While interpretable, this may compress nuance. Future work should explore alternative thresholds and compare against continuous or composite outcomes (e.g., conversational persistence combined with satisfaction). Furthermore, some GLM coefficients were not estimable because certain arm-context combinations were rarely observed. Longer deployments, if possible, could reduce this sparsity. Finally, our evidence is majorly based on descriptive summaries and GLMs. Stronger claims may benefit from off-policy evaluation and comparisons against established baselines.

References

[1] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. 2018. Hindsight Experience Replay. arXiv:1707.01495 [cs.LG] https://arxiv.org/abs/1707.01495

[2] Rohan Banerjee, Rajat Kumar Jenamani, Sidharth Vasudev, Amal Nanavati, Katherine Dimitropoulou, Sarah Dean, and Tapomayukh Bhattacharjee. 2025. To ask or not to ask: Human-in-the-loop contextual bandits with applications in robot-assisted feeding. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1378–1384.

[3] Kim Baraka and Manuela Veloso. 2015. Adaptive Interaction of Persistent Robots to User Temporal Preferences. In *Social Robotics*, Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi (Eds.). Springer International Publishing, Cham, 61–71.

[4] William Cai, Josh Grossman, Zhiyuan Jerry Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph Jay Williams, and Sharad Goel. 2021. Bandit algorithms to personalize educational chatbots. *Machine Learning* 110, 9 (2021), 2389–2418.

[5] Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan. 2019. The assistive multi-armed bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 354–363.

[6] Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. *Advances in neural information processing systems* 24 (2011).

[7] Yuhang Che, Allison M. Okamura, and Dorsa Sadigh. 2020. Efficient and Trustworthy Social Navigation via Explicit and Implicit Robot–Human Communication. *IEEE Transactions on Robotics* 36, 3 (June 2020), 692–707. doi:10.1109/tro.2020.2964824

[8] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. 2020. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 181–190.

[9] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741 [stat.ML] https://arxiv.org/abs/1706.03741

[10] Marlena R. Fraune, Selma Šabanović, and Takayuki Kanda. 2019. Human Group Presence, Group Characteristics, and Group Norms Affect Human-Robot Interaction in Naturalistic Settings. *Frontiers in Robotics and AI* Volume 6 - 2019 (2019). doi:10.3389/frobt.2019.00048

[11] Yuan Gao, Wolmet Barendregt, Mohammad Obaid, and Ginevra Castellano. 2018. When robot personalisation does not help: Insights from a robot-supported learning study. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 705–712.

[12] Koji Inoue, Yuki Okafuji, Jun Baba, Yoshiki Ohira, Katsuya Hyodo, and Tatsuya Kawahara. 2025. A Noise-Robust Turn-Taking System for Real-World Dialogue Robots: A Field Experiment. arXiv:2503.06241 [cs.RO] https://arxiv.org/abs/2503.06241

[13] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. 2012. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*. Springer, 199–213.

[14] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*. 2938–2946.

[15] Chanwoo Kim, Joonhyeok Lee, Eunwoo Kim, and Kyungjae Lee. 2024. Time-Varying Preference Bandits for Robot Behavior Personalization. *APPLIED SCIENCES-BASEL* 14, 23 (2024).

[16] Jens Kober, J. Andrew Bagnell, and Jan Peters. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* 32, 11 (2013), 1238–1274. doi:10.1177/0278364913495721

[17] Jens Kober and Jan Peters. 2014. *Reinforcement Learning in Robotics: A Survey*. Springer International Publishing, Cham, 9–67. doi:10.1007/978-3-319-03194-1_2

[18] Amy Koike, Yuki Okafuji, Kenya Hoshimure, and Jun Baba. 2025. What drives you to interact?: The role of user motivation for a robot in the wild. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 183–192.

[19] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press. doi:10.1017/9781108571401

[20] Guiying Liu, Tatjana Korbanka, Jasmin Timm, Naomi Veit, Laura Maier, Markus Huff, and Frank Papenmeier. 2025. The Bystander Effect in Human-Robot Interaction: How the Presence of Bystanders Inhibits Help for a Social Robot. (Sep. 2025). doi:10.23668/psycharchives.21242

[21] Marcos Maroto-Gómez, María Malfaz, José Carlos Castillo, Álvaro Castro-González, and Miguel Ángel Salichs. 2024. Personalizing activity selection in assistive social robots from explicit and implicit user feedback. *International Journal of Social Robotics* (2024), 1–19.

[22] David Issa Mattos, Jan Bosch, and Helena Holmström Olsson. 2019. Multi-armed bandits in the wild: Pitfalls and strategies in online experiments. *Information and Software Technology* 113 (2019), 68–81.

[23] Zi Haur Pang, Yahui Fu, Divesh Lala, Mikey Elmers, Koji Inoue, and Tatsuya Kawahara. 2025. Does the Appearance of Autonomous Conversational Robots Affect User Spoken Behaviors in Real-World Conference Interactions?. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, Article 198, 8 pages. doi:10.1145/3706599.3720179

[24] Ahmed Hussain Qureshi, Yutaka Nakamura, Yuichiro Yoshikawa, and Hiroshi Ishiguro. 2016. Robot gains social intelligence through multimodal deep reinforcement learning. In *2016 IEEE-RAS 16th international conference on humanoid robots (humanoids)*. IEEE, 745–751.

[25] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Ilhan Aslan, and Elisabeth André. 2018. Drink-o-mender: An adaptive robotic drink adviser. In *Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction*. 1–8.

[26] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Stefan Wagner, and Elisabeth André. 2019. Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback. In *Proceedings of the 12th ACM international conference on PErvasive technologies related to assistive environments*. 247–255.

[27] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Stefan Wagner, and Elisabeth André. 2019. Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Rhodes, Greece) *(PETRA '19)*. Association for Computing Machinery, New York, NY, USA, 247–255. doi:10.1145/3316782.3316791

[28] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. 2018. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning* 11, 1 (2018), 1–96.

[29] Sebastian Schneider and Franz Kummert. 2017. Exploring embodiment and dueling bandit learning for preference adaptation in human-robot interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1325–1331.

[30] Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. 2020. Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives. *Frontiers in Robotics and AI* Volume 7 (2020). doi:10.3389/frobt.2020.532279

[31] William R Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.

[32] Konstantinos Tsiakas, Maher Abujelala, and Fillia Makedon. 2018. Task Engagement as Personalization Feedback for Socially-Assistive Robots and Cognitive Training. *Technologies* 6, 2 (2018). https://www.mdpi.com/2227-7080/6/2/49

[33] Astrid Weiss, Judith Igelsböck, Manfred Tscheligi, Andrea Bauer, Kolja Kühnlenz, Dirk Wollherr, and Martin Buss. 2010. Robots asking for directions: the willingness of passers-by to support robots. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (Osaka, Japan) *(HRI '10)*. IEEE Press, 23–30.

[34] David Whitney, Eric Rosen, James MacGlashan, Lawson L. S. Wong, and Stefanie Tellex. 2017. Reducing errors in object-fetching interactions through social feedback. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 1006–1013. doi:10.1109/ICRA.2017.7989121

Received 2025-09-30; accepted 2025-12-01