IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 August 2025, accepted 25 August 2025, date of publication 8 September 2025, date of current version 15 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3607165

RESEARCH ARTICLE

A Systems-Theoretic Approach to Mental State Estimation for Theory-of-Mind-Aware Social Robots

MARIA L. MORÃO PATRÍCIO AND ANAHITA JAMSHIDNEJAD
Department of Control and Operations, Delft University of Technology, 2629 HS Delft, The Netherlands

Corresponding author: Maria L. Morão Patrício (M.L.MoraoPatricio@tudelft.nl)

This work was supported by Delft University of Technology (TU Delft) AI Laboratories and Talent Program.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Human Research Ethics Committee of TU Delft under Approval No. 3780.

ABSTRACT Social robots are increasingly deployed in fields such as health care and education to support users through social interactions. Nonetheless, these robots mostly rely on black-box machine learning methods that lack awareness of the mental states of their users, which often leads to unnatural behavior. To address this, we propose three model-based techniques for real-time estimation of invisible mental states of humans. Each method adapts the extended Kalman filter and incorporates a validated dynamic model of human mental states. These mental state estimators are designed for human-robot social interactions and personalize their parameters using initial user data. When tested with 10 human participants interacting with a NAO robot, the mental state estimators reduced the average error in estimation and prediction of mental states across all participants by 3% (i.e., from 12% to 9%), with improvements of up to 13% for individual participants. These results demonstrate the potential of integrating such state estimators into the behavioral control systems of social robots to enhance their awareness of the mental states of users.

INDEX TERMS Mental state estimation, theory-of-mind-aware social robots, mathematical model of mind, human–robot social interaction, systems theory for social robotics.

I. INTRODUCTION

In recent years, social robots (SRs) have increasingly been deployed to assist, care for, and entertain users across a wide range of applications. These include deployment in education for tutoring purposes [1], [2], enhancing social and cognitive skills for children with autism [3], [4], [5], administering post-stroke rehabilitation [6], [7], distracting patients during pediatric medical interventions [8], and providing care for older adults and individuals with cognitive disabilities [9], [10], [11], [12].

Despite the promise of these applications in enhancing the quality of life of users, the systems that govern the behavior of SRs are often still perceived as rudimentary and limited [13], [14]. Most SRs rely on black-box machine learning approaches to control their behavioral responses according to metrics related to directly measurable performance of their users, without awareness of their mental states. For example, [15] and [16] applied deep neural networks to learn proper control policies for SRs directly from human demonstrations. Furthermore, [17] and [18] employed Deep Reinforcement Learning for optimal action selection in social robots. While [17] and [18] incorporated some modeled elements — such as biological functions or social norms — into the input or reward function, the main control approach was the machine learning module, and it did not account for the internal mental states of users. Only recently have a few projects started incorporating cognitive states of the users, such as emotions, beliefs, and goals, into behavioral control of SRs [8], [19], [20], or to attribute such states to SRs by embedding cognitive models directly into their behavioral steering systems [21].

Meaningful interactions require that SRs respond to users, reflecting an understanding of their mental states [13], [19],

The associate editor coordinating the review of this manuscript and approving it for publication was Yangming Lee.

[22], [23]. This, in turn, requires accurate modeling and real-time estimation of the mental states of users [13], [19], [23]. A promising approach is to model the dynamical evolution of mental states, based on the inputs (i.e., stimuli perceived by users) and outputs (i.e., actions performed by users), as proposed in [24]. This model, Mathematical Model of Mind (MMM), which has been developed in [24] and subsequently extended for human-robot social interactions in [20], can be used to directly estimate the mental states of users, as illustrated in Figure 1a. However, when measurements are scarce or inaccurate, the estimation based on MMM alone is prone to errors [20].

To enhance the estimation of human mental states, we propose embedding the MMM within a state estimator, as illustrated in Figure 1b. This approach is motivated by a known limitation of the standalone MMM: in the absence of frequent measurements, prediction errors tend to accumulate over time, degrading the quality of the mental state estimates. This limitation is common among open-loop predictive models operating in dynamic, uncertain environments, and does not reflect a fundamental flaw in MMM. Consequently, the issue cannot be fully resolved by replacing the model. Moreover, as demonstrated in [24], MMM achieved an average mental state estimation accuracy of 81.55%, making it a highly promising candidate. Additionally, it is, to our knowledge, the only dynamic cognitive model formulated within a state space framework. This motivates us to retain the model, while introducing additional mechanisms to mitigate the observed accumulation of estimation error. Embedding the model within a state estimator enables us to refine predictions based on available measurements, thereby maintaining higher estimation accuracy over time [25].

Improved accuracy in mental state estimation enables the control systems of SRs to respond more effectively to the varying needs of their users [13], [22]. This, in turn, promotes natural, human-like interactions between SRs and their users.

The remainder of this paper is organized as follows: We describe the main contributions of this paper in Section I-A. Section I-B reviews relevant work on model-based control of SRs that forms the foundation for our approach, as well as existing state estimation techniques relevant to the purposes of this paper. Section II describes the adaptations and enhancements made to tailor existing state estimation methods to the domain of human-robot social interactions. The experimental setup and assumptions are detailed in Section III. Section IV presents and discusses the main results of the paper. Finally, Section V concludes the paper and gives directions for future research.

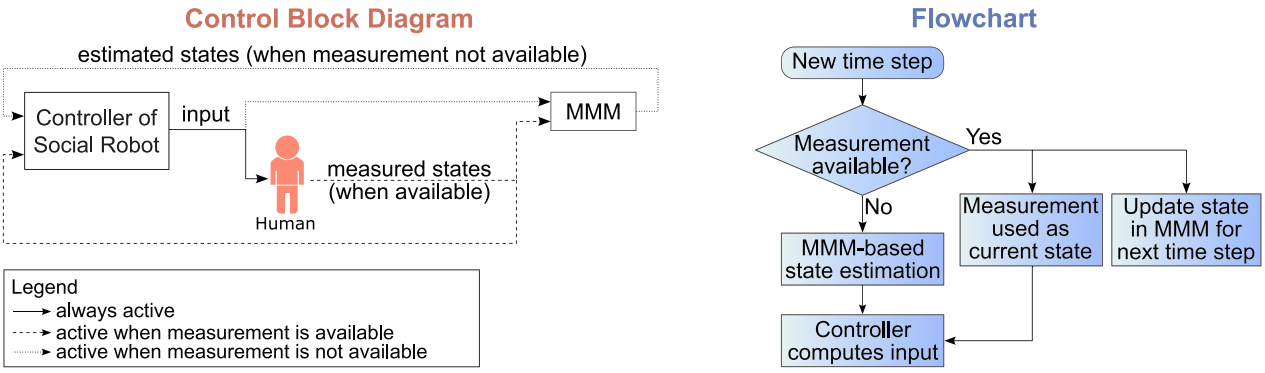
A. MAIN CONTRIBUTIONS

This paper presents three main contributions that advance Human-Social-Robot Interaction (HSRI) beyond previous work, such as [20], in which MMM was used in an open-loop estimation configuration (i.e., mental states were predicted solely based on model dynamics, without correction from measurements), leading to error accumulation over time. Core contributions of this paper are outlined next:

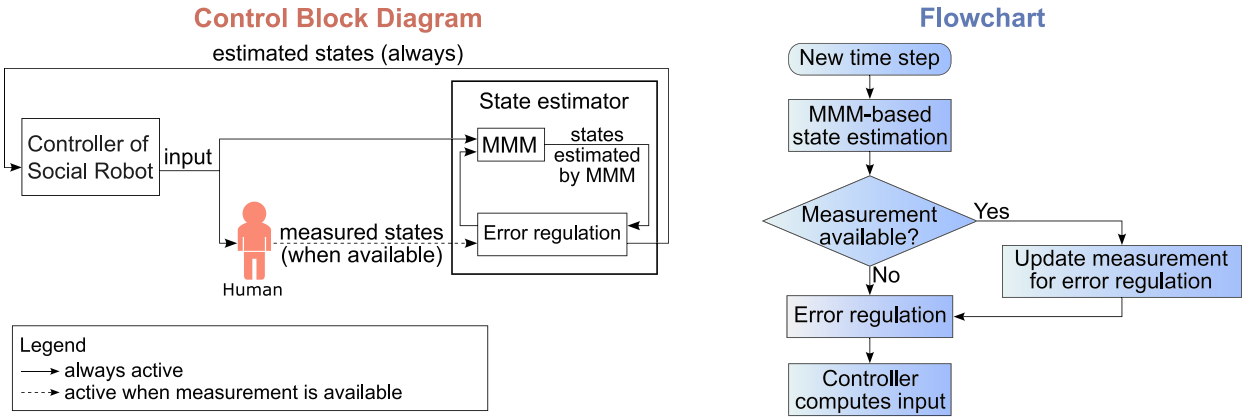
- This paper presents, to our knowledge, the first deployment of systems-theoretic, model-based state estimators — specifically extended Kalman filter (EKF) variants — for estimation of invisible mental states of humans in HSRIs. By adapting state estimation techniques (traditionally used solely for physical and engineering systems) to abstract, dynamic cognitive states, our work bridges a gap between systems and control theory and social robotics. Key technical adaptations include the construction of personalized noise covariance matrices and the modification of EKF computations to handle the challenge of sparse measurements inherent in HSRIs, resulting in three EKF-based estimator variants. This model-based approach offers interpretable, real-time mental state tracking, in contrast with black-box and heuristic methods commonly deployed in HSRIs that typically lack dynamic tracking capabilities.
- We address the degradation problem identified in earlier deployments of MMM [20], in which estimation accuracy declines with sparse user feedback. While [24] introduced MMM as a theoretical framework for mental state estimation and [20] applied this model within HSRIs without a state estimator (see Figure 1a), our work closes this gap by embedding MMM within a personalized, model-based state estimation framework (see Figure 1b).
- We evaluate whether a state estimator based on MMM can improve the accuracy of the mental state estimations for users in HSRIs. Using data collected from 10 human participants interacting with a social robot [20], we evaluate how the proposed EKF-based estimators can potentially improve mental state predictions. Our results show that, compared to the baseline MMM estimator used in [20], the EKF-based estimators reduce prediction errors by an average of 3% across participants, with improvements of up to 13% for individual cases. This is a crucial step towards improving the efficacy of model-based controllers that steer the behavior of SRs in HSRIs based on the evolution of mental states of their users, as previously proposed in [20].

B. RELATED WORK

In [24], a model — referred to as the MMM — is proposed for SRs to represent the dynamic, interconnected processes of human perception, cognition, and decision-making. The MMM is formulated using an input-output state space representation, which enables its integration with established systems and control techniques to enhance HSRIs. In particular, the MMM can be embedded within a model-based controller, enabling SRs to autonomously select actions that optimize users' mental states (e.g., minimizing anxiety or maximizing engagement). To do so, accurate estimation of mental states of the users in real time is



(a) Control block diagram and flowchart representing the system without a state estimator, as proposed in [20]. When a measurement is not available, the MMM computes an estimation of the current state based on the control input and previous state. Then, this estimation is fed to the controller. When a measurement is available, that measurement is fed to the controller of the SR as the current state, and it is used to update the state of the MMM for the next time step.



(b) Control block diagram and flowchart representing the system with a state estimator. The state estimator begins by predicting the system state using the MMM. This prediction is then refined through an error regulation process that merges the model-based estimate with available measurements to minimize the estimation error. When a new measurement becomes available, it is incorporated into this regulation step. The resulting estimated current state is then provided to the controller of the SR.

FIGURE 1. Comparison of the control architectures proposed in [20] and in this paper, illustrating how the current state is determined for use by the predictive controller. Each subfigure represents one architecture through a control block diagram and a flowchart.

essential for SRs. Thus, MMM can also be incorporated into state estimators — mathematical tools for inferring internal state variables of a dynamic system, based on available observations/measurements and a model of the system — to enhance the inference of SRs about the beliefs, goals, and emotions of users.

In [20], MMM was embedded within a predictive model-based controller that steered the behavior of a NAO robot during interactions with users who were solving chess puzzles. The robot dynamically selected the difficulty of the puzzles and chose whether to display entertaining behaviors to optimize the engagement and frustration of participants, and to stimulate prolonged interactions. Parallel to its predictive role, MMM also estimated the evolution of the mental states (i.e., beliefs, goals, emotions) of participants throughout the interaction session. Due to potential errors in these estimations over time, participants were occasionally asked to self-report their beliefs, goals, and emotions. To minimize disruption to participants, however, these self-report occasions were limited to, on average, 10 reports per 35 minutes (roughly one measurement every 3.5 minutes).

Although the proposed approach outperformed a conventional rule-based controller, it was observed that, for some participants, the errors by MMM in estimating the beliefs, goals, and emotions increased significantly between consecutive measurements. Therefore, the performance of the controller degraded over time in the absence of new measurements, leading to sub-optimal behavior by the robot in that interval. For these participants, increasing the accuracy of the estimation of mental states is critical to maintaining a consistently high-quality interaction. To tackle this issue, we propose model-based state estimation methods that integrate the previously validated MMM [20], [24] to track the beliefs, goals, and emotions of users.

Among state-of-the-art state estimation methods, the Kalman filter (KF) remains one of the most widely used

algorithms, mainly due to its computational efficiency and its optimal state estimation in linear systems affected by Gaussian noise [26], [27]. Although the KF is limited to linear systems, the EKF is an adaptation of this algorithm for non-linear models [26], [27]. Other alternatives, such as the unscented KF and the particle filter, are widely used in highly non-linear systems, especially when, due to its linearization, the EKF introduces significant approximation errors [27]. Nonetheless, these approaches come at the cost of substantially increased computational demands, compared to EKF. In our case, while the MMM is a nonlinear model, the degree of nonlinearity is not severe enough to justify the use of more advanced filters [27] (see Section II-B1 for model description). EKFs capture the necessary dynamics with sufficient accuracy, without incurring the computational complexity of unscented KF or particle filtering — an essential requirement in HSRIs. Furthermore, while model-free estimators, such as deep neural networks, can be used for state estimation, our goal is to leverage a validated and interpretable model of human mental state dynamics. Additionally, training such model-free approaches typically requires large datasets, which are generally unavailable in HSRIs. For these reasons, we focus on model-based estimation techniques.

All in all, the EKF offers a suitable balance between performance and computational cost. Consequently, it was selected for integration with MMM to estimate the mental states of humans in HSRIs.

II. METHODOLOGY

In this section, we describe the main steps required to develop and implement a model-based state estimator for estimating the invisible mental states of humans, enhancing various human-centered applications, particularly HSRIs. In Section II-A, we detail how the process and measurement noise covariance matrices, which are essential tuning parameters of KFs, are constructed for a state estimator in the context of MMM [24]. In Section II-B, we explain the procedure for adapting EKFs to estimate the mental states of humans, while dealing with model uncertainties and scarce measurements. We specifically present three versions of EKF developed for this purpose.

A. CONSTRUCTION OF THE NOISE COVARIANCE MATRICES

In order to deploy an EKF, the process noise covariance matrix Q and the measurement noise covariance matrix R, including key parameters of KFs [26], [28], must be defined. A suitable choice of these matrices is crucial for a satisfactory performance of KFs [29], [30]. We note that the construction of the noise covariance matrices Q and R is not presented as a general-purpose methodological contribution. Rather, it is a domain-specific adaptation developed to support the use of EKF in HSRIs, where mental state measurements are subjective, sparse, and limited in number. The following discussion motivates such an adaptation by reviewing existing methods and demonstrating their incompatibility with the HSRI settings.

Although various approaches have been proposed in the literature to estimate these matrices from data (see, e.g., [30], [31], [32], [33], [34]), these methods mostly suit physical systems, i.e., systems that follow laws of physics, and thus, the state estimator should track a quantified physics-based variable as the system state. For example, Mehra [31] developed an adaptive approach based on performing a statistical analysis of the KF to improve the estimation of the noise covariance matrices. This algorithm can be used when initial estimates of the two matrices are available, but the KF that results from applying these matrices is not optimal (i.e., the KF does not provide the minimum mean-squared error estimate of the states). They demonstrated a practical implementation of their approach in an inertial navigation system. Also in the context of navigation, Wu et al. [30] implemented a learning-based adaptive approach to estimate the measurement and process noise covariance matrices for a KF that integrated the signals from a global navigation satellite system with signals from an inertial navigation system. Furthermore, Feng et al. [33] proposed a recursive algorithm for estimating the process noise covariance matrix in real time and demonstrated its effectiveness in a simulation of a simple positioning system that modeled position, velocity, and acceleration.

Given that these approaches have been developed for physical systems, they are based on assumptions or have characteristics that make them unsuitable for EKFs designed to estimate the mental states of humans. Most state-of-the-art approaches rely on statistical analyses of the signals and measurements retrieved from the system [26]. For the sake of accuracy, these approaches require a large number of data samples [30], [31], [32], [34], whereas in HSRI contexts, collecting a sufficiently large number of samples for mental states from humans is not feasible. Moreover, some of these approaches assume linear models (see, e.g., [32], [33]), whereas MMM requires nonlinear functions to accurately capture the complex processes of perception, cognition, and decision-making of humans [20], [24].

Apart from the initial offline construction of the noise covariance matrices, it is also possible to adapt those matrices online to account for errors in the initial estimation. However, such approaches typically require a large amount of data. For example, the adaptive KF proposed in [31] used 10 batches of 950 points in their experiments, and explicitly stated the need for large amounts of data to ensure good performance. Odelson et al. [34] presented another statistics-based adaptive approach that improved upon [31] to estimate Q and R matrices online. In this approach, the accuracy increases with data size, requiring several thousand samples for reliable estimation. Authors in [30] adapted the noise covariance matrices online using neural networks that adjust the noise characteristics in real time. The amount of samples used for training is around 30 times larger than that in [31]. Feng et al. [33] proposed a recursive adaptive approach

that can estimate Q with significantly fewer samples. This, however, is realized at the expense of certain requirements — including linearity and very accurate knowledge of system dynamics and measurement noise — that make this approach impractical for most HSRI settings. Hence, similarly to the methods used to construct the initial values of Q and R, the adaptive approaches in the literature are well-suited for physical systems with high-frequency and abundant measurements, but are not applicable in settings that involve abstract internal mental states and infrequent, subjective measurements. Therefore, relying on state-of-the-art adaptive approaches for estimating or correcting Q and R online is also not feasible in our domain.

Finally, another approach commonly followed in practice is to generate an initial estimate of the two noise covariance matrices based on intuition and fine-tune the parameters based on performance when running the EKF [25]. Therefore, in order to construct the noise covariance matrices for EKFs that estimate the mental states of humans, it is necessary to delve into the physical meaning of the elements of these matrices. This allows the creation of a parallel between the typical systems for which these approaches have been developed and MMM [24], so we can adapt the design of the EKF accordingly. In the context of human cognition, the measurements for estimating the mental states (beliefs, goals, emotions) may be taken in two ways: (i) using physical sensors that indirectly measure psychological states of users (e.g., measuring physiological signals, such as heart rate, to infer the anxiety levels of a user, or using visual recognition to quantify user engagement); (ii) asking users to directly report their beliefs, goals, and emotions. On the one hand, the first approach is less intrusive and less disruptive in the context of HSRIs. On the other hand, because these measurements are indirect indicators of mental states, they may not accurately reflect the true mental states estimated by MMM, which can lead to increased errors in the estimation process. Moreover, not all relevant mental states are always retrievable from such measurements. Therefore, Morão Patrício and Jamshidnejad [20] applied the second approach (i.e., directly receiving feedback about the mental states from users) to measure the mental states of users during experimental HSRIs, where data gathered in those experiments was used in this paper to tune and validate the developed state estimators.

To construct the two matrices, which are personalized per participant, we use the sequence of data points $\mathcal{D}$ collected in two HSRIs that were part of the experiment described in [20] (the dataset collected in these experiments can be found in [35]). On average, 41 data points were collected per participant per session. Each data point contains the values of the input vector $\boldsymbol{u}(k)$ (i.e., environmental data related to the interaction that was accessible in the real world) and the measurement of the state vector $\boldsymbol{x}(k)$ (i.e., the beliefs, goals, and emotions of each participant) at time step k. Data is stored in sequence $\mathcal{D}$ in order of collection, i.e., based on time step k.

1) PROCESS NOISE COVARIANCE MATRIX

The process noise covariance matrix Q reflects the uncertainty of the estimations made by the model in a KF about the values of the dynamic state variables [26], [27]. In order to construct the Q matrix for a KF that estimates mental states of humans, we assume that the noise on the different variables is uncorrelated, a common assumption in literature when constructing this matrix [26]. The assumption implies that matrix Q is diagonal and that each diagonal element $Q_{i,i}$ reflects the uncertainty of MMM when estimating the time evolution of mental state $x_i(k)$. To determine this uncertainty, we evaluate the prediction error of MMM per variable $x_i(k)$ using data from real-life HSRI experiments. Specifically, we estimate each element $Q_{i,i}$ by calculating the mean squared error between the prediction made by the identified MMM for the value of mental state $x_i(k)$ and the corresponding value from the collected data. Therefore, Q is given by:

$$Q = \text{diag}(\sigma_1^{\text{Q}}, \ldots, \sigma_n^{\text{Q}}) \tag{1}$$

where n is the number of mental states (i.e., the cumulative number of possible beliefs, goals, and emotions) and σ_i^{Q} for $i = 1, \ldots, n$ is determined via:

$$\sigma_i^{\text{Q}} = \frac{1}{|\mathcal{D}|} \sum_{k \in \{1,\ldots,|\mathcal{D}|\}} (\hat{x}_i^{\text{MMM}}(k) - x_i^{\text{m}}(k))^2 \tag{2}$$

where $\mathcal{D}$ is the sequence of data gathered from real-life experiments and stored in chronological order, $\hat{x}_i^{\text{MMM}}(k)$ is the value that is estimated by MMM for mental state $x_i(k)$ for time step k, and $x_i^{\text{m}}(k)$ is the measured value of the variable (captured from real-life experiments for mental state $x_i(k)$). In (2) all values measured for the i^{th} mental state at all time steps included in the experimental dataset $\mathcal{D}$ are used to determine σ_i^{Q}. Moreover, $|\cdot|$ represents the cardinality of the sequence.

2) MEASUREMENT NOISE COVARIANCE MATRIX

The measurement noise covariance matrix R represents the noise in the system measurements [26]. In physical systems, the system measurements are often obtained via sensors. Apart from using data-driven approaches, such as those discussed earlier in this section, the measurement noise covariance matrix is often extracted based on the characteristics of the sensors or via statistical analyses of the measurements obtained from the system [25], [28].

In the current context, since the measurements were provided directly by each user (rather than collected through physical sensors), it is not possible to construct matrix R based on sensor information, as is common for this matrix. Moreover, since the number of samples captured during the experimental HSRIs is limited, it is not possible either to conduct complex statistical analyses on the data to deduce meaningful patterns on the characteristics and evolution of the data, which could assist in constructing R.

Therefore, an alternative approach was adopted to construct the measurement noise covariance matrices for

each user. Specifically, the elements of the R matrix were estimated using state measurements collected under steady-state conditions, for which it is necessary to keep the inputs to the system constant. Under such conditions, any variation observed in the measured states is assumed to arise from measurement noise rather than changes in the underlying system. Accordingly, the variance of these measurements was used to estimate the measurement noise covariance.

Given the aforementioned limitations in data collection inherent to experiments with humans, it was not straightforward to re-create such steady-state conditions while collecting measurements for two reasons. First, prolonged constant conditions may lead to distraction or boredom of participants, which increases the measurement noise. Second, due to the scarcity of the measured data, dedicating many measurements solely to identifying the noise covariance matrix is inefficient. Therefore, the data used to identify MMM was reused to identify the noise covariance matrix R, by isolating sequences of at least three consecutive data points with similar values for the inputs. As mentioned earlier in this section, each data point includes the vector of all mental states in a time step. Sequences with fewer than three data points were excluded because they provided too little data to reliably estimate the variability caused by measurement noise. The threshold of three points was chosen as a practical compromise, balancing the need for reliability with the limited amount of available data. The variance of the i^{th} measured mental state per sequence $\mathcal{S}_\ell$ is computed by:

$$\sigma^{\text{R}}_{i,\ell} = \frac{1}{|\mathcal{S}_\ell|} \sum_{x^{\text{m}}_i(k)\in\mathcal{S}_\ell} \left(x^{\text{m}}_i(k) - \bar{x}_{i,\ell}\right)^2 \tag{3}$$

where $\bar{x}_{i,\ell}$ is the average value of $x^{\text{m}}_i(k)$ across all time steps within sequence ℓ. Assuming a total number n^{seq} of such isolated sequences, for each mental state, the variance $\sigma^{\text{R}}_{i,\ell}$ is averaged across all the sequences, yielding a representative value of the variance for that mental state. We have:

$$\sigma^{\text{R}}_i = \frac{1}{n^{\text{seq}}} \sum_{\ell\in\{1,\dots,n^{\text{seq}}\}} \sigma^{\text{R}}_{i,\ell} \tag{4}$$

As it is common in practical implementations of KFs, we assume that the measurement noise covariance matrix R is diagonal [28]. Thus, the R matrix is given by:

$$R = \text{diag}(\sigma^{\text{R}}_1, \dots, \sigma^{\text{R}}_n) \tag{5}$$

B. IMPLEMENTATION OF THE EXTENDED KALMAN FILTER

Given the central role of the process and measurement models in the EKF, we begin this section by presenting their general form, followed by an overview of the classical EKF, and a description of the three EKF variants proposed in this work.

1) PROCESS AND MEASUREMENT MODELS

In our implementation, the state transition function $f(\cdot)$ of the process model used in the EKF is derived from the MMM, originally presented in [24] and subsequently refined in [20]. The state transition function captures the dynamic evolution of three groups of mental state variables — beliefs, goals, and emotions — where each category can include multiple variables. The measurement model links these internal states to user self-reports.

The general formulations of the process and measurement models are, respectively, given by:

$$\boldsymbol{x}(k) = f\left(\boldsymbol{x}(k-1), \boldsymbol{u}(k-1)\right) + \boldsymbol{w}(k-1) \tag{6a}$$

$$\boldsymbol{x}^{\text{m}}(k) = h\left(\boldsymbol{x}(k)\right) + \boldsymbol{v}(k) \tag{6b}$$

where $\boldsymbol{x}(k)$ is the vector of all mental states at time step k, and $\boldsymbol{x}^{\text{m}}(k)$ is the vector of all measured (e.g., self-reported) mental states at time step k. The vector $\boldsymbol{u}(k)$ includes the inputs at time step k, i.e., the real-world data that affects the cognition of the user. The vectors $\boldsymbol{w}(k)$ and $\boldsymbol{v}(k)$ are, respectively, the process and measurement noise, and are assumed to follow a Gaussian distribution with zero mean and Q and R covariance matrices, respectively. The function $f(\cdot)$ is the state transition function of MMM [20], [24], and $h(\cdot)$ is the measurement function. Both $f(\cdot)$ and $h(\cdot)$ are, in general, vector functions.

Based on the formulation in [20], we briefly summarize the structure of the state transition function $f(\cdot)$, focusing on its dependence on the state variables. This is the portion relevant for the linearization step in the EKF, which requires computing the Jacobian with respect to $\boldsymbol{x}(k)$ (explained in detail in Section II-B2 and employed via (9)). For a detailed description of the full model, we refer the reader to [20]. The following equation presents $f_i(\cdot)$, the i^{th} component of the nonlinear state transition function $f(\cdot)$, corresponding to the update of the i^{th} state variable x_i:

$$\begin{aligned} f_i\left(\boldsymbol{x}(k-1), \boldsymbol{u}(k-1)\right) &= w_i x_i(k-1) \\ &+ \sum_{i\neq j} w_{ji}(k) x_j(k-1) + f^{\text{input}}(\boldsymbol{u}(k-1)) \end{aligned} \tag{7}$$

where w_i is a scalar weight that regulates the influence of $x_i(k)$ on its next realization. The terms $w_{ji}(k)$ are weights that represent the influence of other mental state variables $x_j(k-1)$, with $j \neq i$, on the evolved value $x_i(k)$ of mental state i. These weights are defined as piecewise constant, depending on the region of the state space. For example, the influence of the emotion "bored" on the goal of "quitting the interaction" may differ depending on whether the emotion is positive (the user is bored) or negative (the user is engaged). In other words, $w_{ji}(k) = w^+_{ji}$ when $x_j(k) > 0$, and $w_{ji}(k) = w^-_{ji}$ when $x_j(k) \leq 0$. This structure results in a globally nonlinear function composed of multiple locally affine components. The function $f^{\text{input}}(\cdot)$ accounts for the influence of external inputs (i.e., real-world data) perceived by the user. However, its detailed structure is omitted here, as it does not contribute to the linearization process used in the EKF.

Additionally, in our application, the measurements directly correspond to the mental states, as they are measured through self-reports from users. Consequently, the measurement function is given as:

$$h\left(\boldsymbol{x}(k)\right) = \boldsymbol{x}(k) \tag{8}$$

2) ADAPTATION OF EXTENDED KALMAN FILTER

Following [26], we divide the estimation made by the EKF into two phases: the prediction phase, represented via (9), and the correction phase, represented via (10).

a: PREDICTION PHASE

The prediction phase is formulated by:

$$\hat{\boldsymbol{x}}^{\mathrm{MMM}}(k) = f\left(\hat{\boldsymbol{x}}(k-1), \boldsymbol{u}(k-1)\right) \tag{9a}$$

$$P^{-}(k) = F(k-1)P(k-1)F^{\top}(k-1) + Q \tag{9b}$$

where $\hat{\boldsymbol{x}}^{\mathrm{MMM}}(k) = [x_1^{-}(k), \ldots, x_n^{-}(k)]^{\top}$ is the vector of all estimated mental states by the MMM at time step k, $f(\cdot)$ is the state transition function given by the state space representation of MMM [24], $F(k)$ is the Jacobian matrix of $f(\cdot)$ evaluated for realized value of the state vector $\hat{\boldsymbol{x}}$ at time step k, and $P(k)$ is the error covariance matrix predicted for time step k.

b: CORRECTION PHASE

The correction phase of the EKF is formulated via:

$$K(k) = P^{-}(k)H^{\top}(k)\left(R + H(k)P^{-}(k)H^{\top}(k)\right)^{-1} \tag{10a}$$

$$\hat{\boldsymbol{x}}(k) = \hat{\boldsymbol{x}}^{\mathrm{MMM}}(k) + K(k)\left(\boldsymbol{x}^{\mathrm{m}}(k) - h\left(\hat{\boldsymbol{x}}^{\mathrm{MMM}}(k)\right)\right) \tag{10b}$$

$$P(k) = (I - K(k)H(k))\,P^{-}(k) \tag{10c}$$

where $P^{-}(k)$ is computed via (9), $h(\cdot)$ is the output function of the state space representation of MMM, $H(k)$ is the Jacobian matrix of $h(\cdot)$ evaluated at $\hat{\boldsymbol{x}}(k)$, $\boldsymbol{x}^{\mathrm{m}}(k)$ is the vector of all measured mental states at time step k, and $P(k)$ is the error covariance matrix.

While the prediction phase, shown in (9), depends only on the model prediction (i.e., on $\hat{\boldsymbol{x}}^{\mathrm{MMM}}(k)$), the correction phase, given by (10), depends also on the measurements (i.e., on $\boldsymbol{x}^{\mathrm{m}}(k)$) captured for the mental states. Therefore, while the prediction phase can be carried out in each discrete time step and is only potentially limited by the computation speed of MMM, the correction phase can only be performed when a measurement of the mental states becomes available.

In practical implementations of MMM, measurements of mental states may be unavailable at every discrete time step, as explained in Section II-A. Therefore, to directly apply both phases of the EKF at each time step k, prior adjustments are required.

We next propose and implement three approaches to circumvent this issue. Figure 2 illustrates the functioning of these three methods.

3) INTERMITTENT EKF

The intermittent EKF employs a common approach for implementing EKFs when measurements are less frequent than the estimation updates. In this approach, the prediction phase is executed at every time step. When measurements are available, the correction phase is also performed, and the estimation is obtained via (10b). In the absence of measurements, the correction phase is skipped, and the state estimate is obtained from (9a), which relies only on the prediction made by MMM (see Figure 2). Figure 3 illustrates the operation of the intermittent EKF, showing when the prediction and correction phases are applied based on the algorithm and the availability of measurements.

Despite its simplicity and common deployment, this approach faces limitations when applied to HSRIs. In such applications, the sparsity of measurements may lead the EKF to rely only on model predictions for most time steps. As a result, it provides limited improvements compared to approaches that do not use a state estimator and estimate the states directly from the model, as in [20]. To overcome this limitation, we propose an alternative approach that more effectively leverages the information gathered from available measurements.

4) FORWARD-FILL EKF

The second approach, called the forward-fill EKF, addresses the limitations of the intermittent EKF by incorporating the measurements more frequently, thereby enabling both phases of the EKF to be executed every discrete time step k. To achieve this, the measurements are forward-filled, i.e., at each time step k, the value of measurement $\boldsymbol{x}^{\mathrm{m}}(k)$ is set to the most recent available measurement. Consequently, if no new measurement is available for q consecutive time steps, the measurement $\boldsymbol{x}^{\mathrm{m}}(k)$ corresponding to time step k is re-used for $\boldsymbol{x}^{\mathrm{m}}(k+1), \ldots, \boldsymbol{x}^{\mathrm{m}}(k+q)$. Figure 4 illustrates the operation of the forward-fill EKF, demonstrating its behavior depending on the availability of measurements.

Although this approach benefits from continuous incorporation of measurement information, it may introduce errors when mental states change frequently or abruptly. Nevertheless, we expect that the contribution of the model to the EKF estimation — combined with the personalized construction of measurement and process noise covariance matrices based on the noise characteristics of the model and measurements per participant — will help to mitigate such measurement errors over time.

5) FORWARD-FILL-EVOLVING EKF

The third and final approach, the forward-fill-evolving EKF, accounts for the decreasing reliability of measurements over time in the absence of new measurements. Since the measurement noise covariance matrix $R(k)$ reflects the expected measurement error, we propose increasing the value of the elements of this matrix during time steps when no new measurements are available. Therefore, matrix $R(k)$ is updated by:

$$R(k) = \begin{cases} R_0, & \text{if mental states are measured at time step } k \\ \alpha\, R(k-1), & \text{otherwise} \end{cases} \tag{11}$$

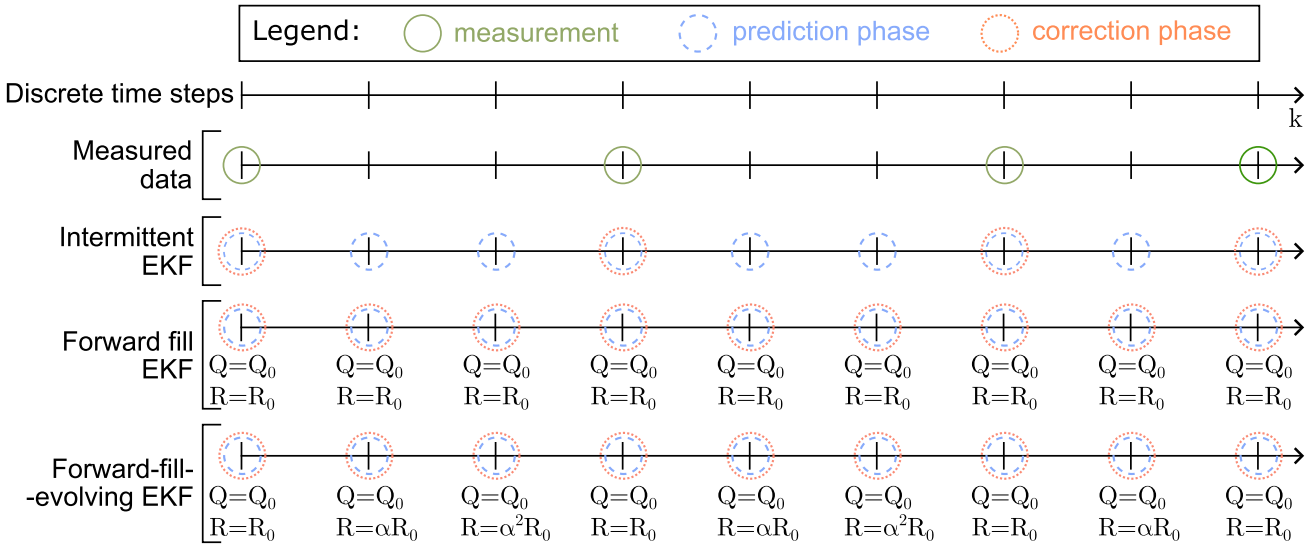


FIGURE 2. Illustration of the functioning of the three proposed versions of the EKF depending on the availability of measurements. The top axis represents the discrete time steps, with green solid circles showing time steps when measurements for mental states are available. The bottom three axes show when the prediction and correction phases are applied for each variant of EKF, depending on measurement availability. Furthermore, the evolution of the Q and R matrices over time is depicted below each axis.

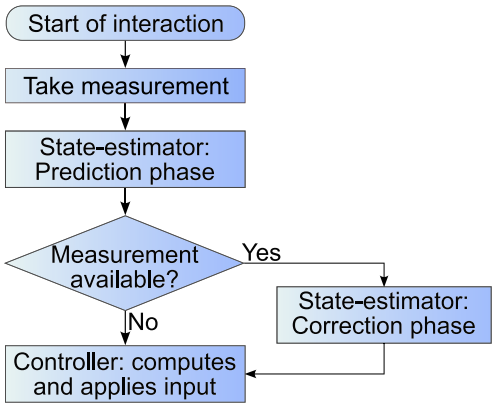


FIGURE 3. Flowchart illustrating the control loop steps for the intermittent EKF. The diagram outlines the sequential steps of this state estimator, depending on the availability of measurements.

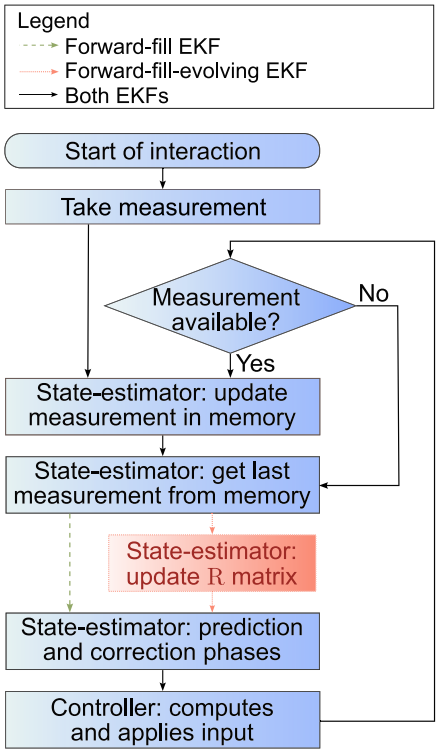


FIGURE 4. Flowchart illustrating the control loop steps for the forward-fill EKF and forward-fill-evolving EKF. The diagram outlines the sequential steps of these two state estimators, depending on the availability of measurements. The additional step specific to the forward-fill-evolving EKF is distinguished with a red dotted line.

where $\alpha > 1$ is a scalar parameter and R_0 is the measurement noise covariance matrix calculated offline. This strategy ensures that whenever a new measurement is unavailable and the previous one is reused, the entries of the $R(k)$ matrix are increased with respect to the previous time step. As a result, compared to a fixed $R(k)$ matrix, the Kalman gain is lower (see (10a)), and the influence of the measurement in the estimation of the updated mental states is reduced (see (10b)). Furthermore, the longer the gap between capturing two measurements, the less the KF relies on the measurements and the more it relies on the model estimation. Figure 2 (bottom plot) depicts how the $R(k)$ matrix evolves with this approach based on measurement availability. The value of parameter α determines the growth of matrix $R(k)$ per time step when there are no new measurements. Thus, α effectively models the rate of decay of the confidence in reused measurements.

This strategy constitutes a simple, domain-specific form of adaptive filtering, as it heuristically modifies the measurement noise covariance $R(k)$ based on the availability of new measurements. Figure 4 illustrates the operation of the forward-fill-evolving EKF, demonstrating its behavior depending on the availability of measurements and how its behavior differs from the forward-fill EKF.

Based on the configuration of these three approaches, we formulate the following three hypotheses:

Hypothesis 1: Intermittent EKF will perform similarly to the estimator in [20].

Hypothesis 2: Forward-fill EKF will outperform the estimator in [20].

Hypothesis 3: Forward-fill-evolving EKF will outperform forward-fill EKF, as it better balances the predictions generated by the MMM with the varying reliability of measurements over time.

These hypotheses are further discussed in Section IV.

6) CONVERGENCE CONSIDERATIONS

Formal convergence guarantees for EKF usually require restrictive assumptions, and EKF-based applications in nonlinear and adaptive settings often rely on empirical assessment of convergence behavior [27], [36], [37], [38]. In the context of this paper, the specific formulation of MMM is customized per participant and HSRI scenario, making a general formal convergence proof for the proposed EKFs infeasible.

Nonetheless, some theoretical insight can be gained by examining the general structure of the state transition function $f(\cdot)$ of MMM. Although $f(\cdot)$ is nonlinear, it is piecewise affine with respect to the state $\boldsymbol{x}(k)$ within local regions of the state space (see (7)). In these regions, the EKF behaves as a standard KF, for which local convergence can be established under classical conditions, such as observability, noise properties, and accurate initial estimates [27], [37]. While these conditions are likely to hold in practice given the structure of the model and the assumptions adopted (see Section II-B1), they should be explicitly verified in each context-specific instantiation of MMM to formally guarantee local convergence. However, even if local convergence is satisfied, it has limited practical use in the present context. During interaction, dynamic cognitive state variables (i.e., beliefs, goals, and emotions) may transition across these affine regions of the state space. As a result, convergence guarantees within those regions do not generalize globally.

Similarly, a general global convergence proof is not tractable, due to the context-specific and personalized nature of cognition modeled through MMM. Each participant-specific model yields unique dynamics, requiring separate convergence analyses. Moreover, any adaptation of the model during deployment — as expected in realistic, long-term HSRIs — invalidates previously established guarantees, making global guarantees impractical in real-world applications.

Given these theoretical limitations, we adopt an empirical approach to assess convergence behavior, since this aligns better with the practical requirements of our setting. In Section IV, we present experimental results that demonstrate how estimation uncertainty evolves by analyzing the trace of the covariance matrix of the proposed state estimators for each participant over time. This provides a practical indication of whether the filters maintain bounded uncertainty over time, supports the safe and effective use of EKFs, and offers an interpretable and adaptive measure for the reliability of estimations in real time during personalized HSRIs.

III. CASE STUDY

In this section, we present the case study used to evaluate the state estimators proposed in this paper. First, we describe the experimental setup in Section III-A, followed by an explanation of the assessment methodology in Section III-B.

A. EXPERIMENTAL SETUP

This experiment assesses the extent to which the proposed model-based mental state estimators improve the online predictions of the mental states of participants, compared to estimations made by MMM without a state estimator (see Figure 1 for a comparison between the two architectures).

The state estimators were evaluated on data we collected within an extensive human-robot interaction study, in which the behavioral control framework of the robot, task design, and system implementation have been detailed in [20]. The dataset includes self-reported mental state measurements from 10 human participants, collected across three sessions each. For completeness, we provide a concise summary of the experimental protocol and robot behavior in this section. Readers interested in the full system design and implementation details are referred to [20].

The experiments were conducted using a NAO V6 humanoid robot [39], a platform widely used in HSRI research due to its versatility, ease of programming, and suitability for social applications [40], with research applications spanning health-care [6], [41] and educational contexts [42]. The high-level behavior control of NAO was implemented in Python using the NAOqi SDK, and was executed on a connected laptop running Windows 10 with an Intel CPU (i7–1185G7 3.0GHz), and 32 GB of RAM. Built-in modules provided by the SDK were employed for speech synthesis and motion control. Motion control also included adapted versions of movements originally designed in Choreographe.

In the HSRI sessions, 10 participants interacted with a NAO Robot [39] while solving chess puzzles (see Figure 5). As outlined in Section I-B, NAO dynamically selected the difficulty level of the puzzles and decided whether to display entertaining behaviors during the interaction, aiming to keep the participants engaged. Meanwhile, NAO evaluated and tracked the mental states, i.e., one value for the belief (related to the difficulty level of the puzzle), two values for the goals (related to quitting the interaction and to skipping the puzzle),

FIGURE 5. Experimental setup of HSRI with NAO robot [39]. Reproduced from [20].

and two values for the emotions (related to the frustration and the boredom) per participant. The objective for NAO was to select actions that minimized the values of goals and emotions, and brought the belief value as close to zero as possible — indicating that the perceived difficulty of the game was suitably balanced, neither too easy nor too hard.

The experiment consisted of three sessions. The first two sessions, each lasting 45 to 60 minutes, were used to collect measurements of the five variables through participant self-reports. On average, 41 measurements were collected per session. The data was then used to identify the parameters of MMM to personalize the model per participant. In the third session, the personalized MMM was used by the robot to track the mental states of the participants. It was also embedded in a predictive model-based controller to select optimal actions for NAO. To mitigate the estimation error accumulation over time that occurs due to a lack of measurements, the participants were requested to provide self-reports of their mental states over time (approximately 10 measurements per participant during the 35-minute interaction).

In our deployment, data from the first two sessions was used to construct the process and the measurement noise covariance matrices Q and R for each participant according to the approach described in Section II-A. The parameter α of the forward-fill-evolving EKF (see (11)) was set to 2. A data-driven tuning approach was not pursued, as it would have required an additional dataset specifically for tuning α after the initial values of the noise covariance matrices Q and $R(k)$ had been obtained. Instead, α was selected as a trade-off between enabling a meaningful comparison between forward-fill EKF and forward-fill-evolving EKF, and maintaining a conservative value (i.e., close to 1) to avoid a substantial increase in $R(k)$. Data from the third session (which we refer to as the validation dataset) was then used as online input to simulate and assess the estimations that each EKF would have produced during the HSRI. These predictions were compared against the predictions done in [20], i.e., the predictions produced by the standalone model. After analyzing the results of the three proposed EKFs, the value of parameter α of the forward-fill-evolving EKF was varied to assess its impact and determine whether the data-driven personalization of this parameter influences the performance of the state estimator.

The intermittent EKF, forward-fill EKF, and forward-fill-evolving EKF were implemented in Python 3.9. The assessment of the EKFs was conducted in an offline setting using a laptop running Windows 10 with an Intel CPU (i7–1185G7 3.0GHz), and 32 GB of RAM.

B. ESTIMATION ASSESSMENT

In order to assess the performance of the three EKF variants proposed in this article and to compare them with the baseline estimations from [20], we used the same validation dataset [35] from that study, treating it as data that is being captured sequentially and online.

Since our goal is to reduce the cumulative prediction error that builds up over time in the absence of frequent measurements, we assessed the accuracy of the estimations made by the EKFs based on the prediction errors that accumulate between consecutive measurements.

To ensure a fair comparison, all estimators were evaluated using the same dataset. The evaluation process starts from a time step k, when a measurement $\boldsymbol{x}^{\mathrm{m}}(k)$ is available (e.g., k_3 in Figure 6). The EKF estimates the state based on the previous estimation $\hat{\boldsymbol{x}}(k-1)$ and the current measurement $\boldsymbol{x}^{\mathrm{m}}(k)$, applying the prediction and correction phases as described in (9) and (10). From this time step until a new measurement is available, for all time steps $\kappa > k$ (e.g., at time steps k_4 and k_5 in Figure 6), the estimations are conducted differently. For the intermittent EKF only the prediction phase is executed, whereas for the forward-fill EKF and forward-fill-evolving EKF, both prediction and correction phases are applied using the last available measurement $\boldsymbol{x}^{\mathrm{m}}(k)$.

When a new measurement becomes available at time step $k+m$ (e.g., at time step k_6 in Figure 6), two operations are performed. First, an estimation is made as if no new measurements were available, as was done for the time step $k+m-1$. The resulting estimation is compared to the actual measurement $\boldsymbol{x}^{\mathrm{m}}(k+m)$. This yields the cumulative prediction error since the last measurement. Second, the measurement $\boldsymbol{x}^{\mathrm{m}}(k+m)$ is incorporated to update the state estimation, which is then used for subsequent predictions.

This process is repeated at each time step when a measurement is available, excluding the first measurement, which serves as initialization. For each participant, the average cumulative error is computed as the mean absolute value of all cumulative errors across their session.

IV. RESULTS AND DISCUSSIONS

In this section, we present and discuss the estimations made by the EKFs for the mental states of the human participants

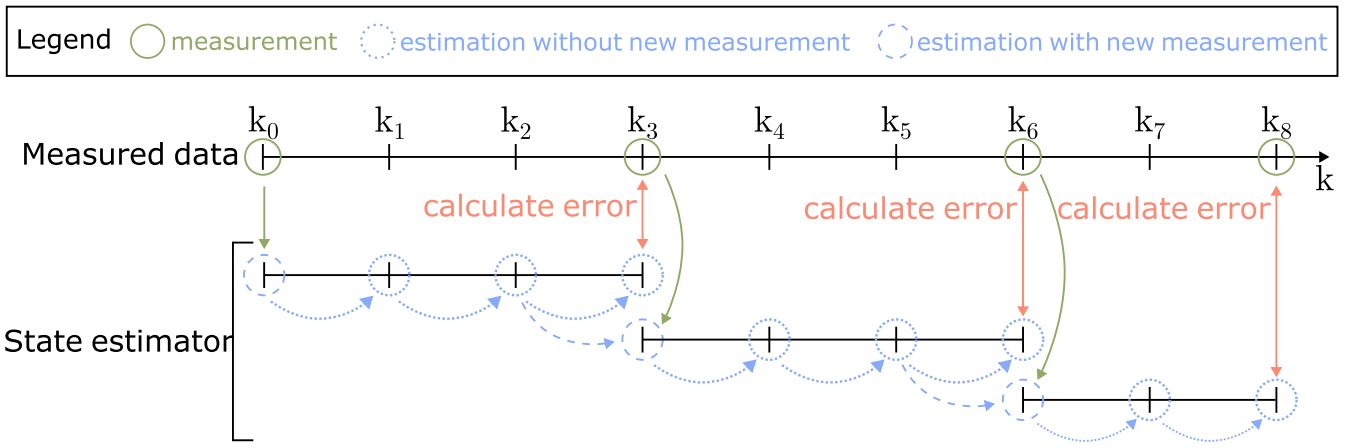


FIGURE 6. Example of computing the cumulative prediction errors between measurements: The measurement at time step k_0 initializes the EKF. For the next time steps k_1 and k_2 with no new measurements, predictions are made using only the model in the case of the intermittent EKF, and using the forward-filled value of the last measurement for the forward-fill EKF and forward-fill-evolving EKF. This is represented by the blue dotted arrows and circles. At time step k_3, a new measurement is available, prompting two parallel computations. First, an estimation is performed assuming no new measurement (blue dotted arrow and circle) that is compared to the actual measurement to compute the cumulative error between time steps k_0 and k_3. Second, an estimation is made for time step k_3 utilizing the measurement $x^{\mathrm{m}}(k_3)$ (specified by the green arrow, where its transition from the previous time step has been shown via a dashed blue arrow). This estimation is then used for future estimations. The same process is repeated at time steps k_6 and k_8, when cumulative errors are also computed. The average cumulative error is given by averaging the cumulative prediction errors at time steps k_3, k_6, and k_8.

who interacted with NAO while solving chess puzzles. In the absence of established state estimators for internal cognitive states in HSRIs, we compare our proposed methods against the standalone open-loop estimator based on MMM used in [20], which represents the current state of practice in this domain. Although advanced state estimation methods exist for physical systems, they are not directly applicable here, due to fundamental differences in measurement quantity and the abstract nature of mental state variables.

A. STATE ESTIMATION ACCURACY

The average cumulative prediction errors between consecutive measurements for each state estimator are presented in Figure 7. For each state estimator and participant, these errors were computed as described in Section III-B. Figure 7 presents the distribution of errors across the ten participants using box plots for each state estimator. The minimum, maximum, median, and mean errors are also provided in Table 1. As expected, the intermittent EKF did not improve prediction accuracy over the baseline results from [20] (see hypothesis 1 in Section II-B). Its performance is nearly identical to that of the original MMM when no state estimator is used (see Figure 7 and Table 1). In contrast, the forward-fill EKF demonstrates a clear prediction enhancement over the baseline where estimations were made by MMM alone. In fact, while the minimum prediction error remains similar, the maximum, median, and mean prediction errors decrease by 2.8%, 4.0%, and 2.9%, respectively. This confirms hypothesis 2.

In hypothesis 3, we theorized that increasing the values of the measurement noise covariance matrix $R(k)$ during periods without new measurements, as implemented in the forward-fill-evolving EKF, would enhance the state estimator performance. However, the initial results did not support this hypothesis. The performance of forward-fill-evolving EKF was comparable to that of forward-fill EKF, as shown in Figure 7 and Table 1. Several factors may explain this outcome: First, the forward-fill EKF already yields significant performance improvements, leaving limited room for further prediction error reduction in this particular HSRI context. In general, mental states of humans are inherently difficult to estimate with high precision, making it unrealistic to expect negligible prediction errors for such states. Second, the mental states captured from the participants in this dataset are relatively stable for most participants, meaning that measurements remain informative and reliable over longer periods. Consequently, penalizing the degradation of measurement accuracy over time by increasing the elements of the $R(k)$ matrix — which distinguishes forward-fill-evolving EKF from forward-fill EKF — may be unnecessary in this context. Finally, the parameter α, which governs how quickly matrix $R(k)$ evolves over time, was set to 2 for all participants, as explained in Section III-A. Since data from the first two sessions was used to estimate the R and Q matrices, and data from the third session was used to assess the performance of the state estimators, a fixed, non-personalized value was used for α, which may explain why the original hypothesis was not supported by the results. The possibility of enhancing the performance of the forward-fill-evolving EKF by personalizing α is, thus, discussed later in this section.

Overall, the modifications offered by the forward-fill EKF proved to be the most effective enhancements over the original estimation method. The results indicate that the

TABLE 1. Relative prediction errors of the mental states for the ten participants across different state estimators. The intermittent EKF, forward-fill EKF, and forward-fill-evolving EKF are three variants developed in this paper, adapted from the standard EKF [27] and whose technical design and implementation are described in Section II-B. These are compared against the "No state estimator" baseline used in [20]. The minimum, maximum, median, and mean values of the average prediction error per participant and state estimator are presented.

State estimator	Minimum error (%)		Maximum error (%)		Median error (%)		Mean error (%)	
	Error	Improvement	Error	Improvement	Error	Improvement	Error	Improvement
No state estimator [20]	4.621	—	20.660	—	11.996	—	11.956	—
Intermittent EKF	4.608	-0.013	20.714	0.054	11.994	-0.002	11.977	0.021
Forward-fill EKF	3.620	-1.001	17.888	-2.772	8.038	-3.958	9.083	-2.873
Forward-fill-evolving EKF	4.290	-0.331	17.564	-3.096	8.091	-3.905	9.322	-2.634

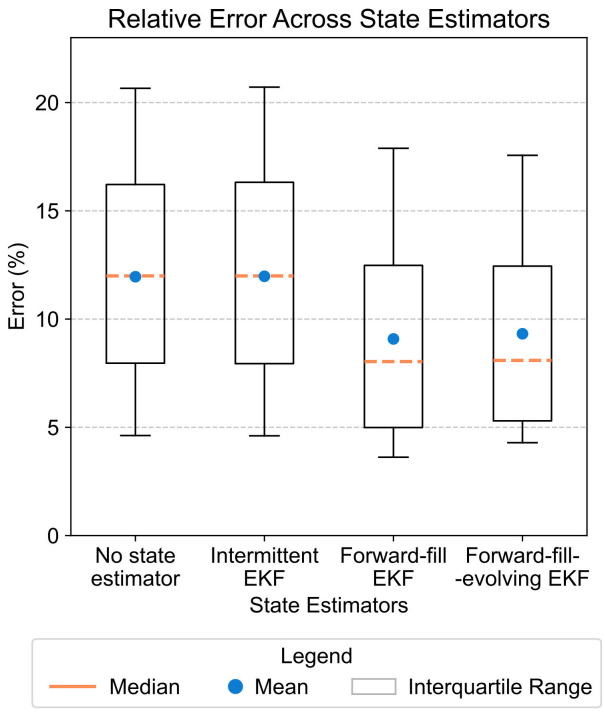


FIGURE 7. Relative prediction error of the mental states of the ten participants across different state estimators. The average prediction errors per state estimator obtained during the interaction with each participant are displayed as box plots. The intermittent EKF, forward-fill EKF, and forward-fill-evolving EKF, which are adapted versions of EKFs [27] developed in this paper and described in Section II-B, are compared with the "No state estimator" baseline used in [20].

average estimation error across all participants is reduced by 2.9% when using the forward-fill EKF. This provides preliminary evidence that the forward-fill EKF enhances estimation accuracy. However, there are considerable differences in the magnitudes of the original estimation errors among the ten participants. Therefore, we further analyze the estimation errors of each participant individually to assess more thoroughly the degree of improvement provided by the proposed EKFs.

To gain more insight into the impact of each state estimator, we analyze the average prediction errors over time for each participant individually. As shown in Figure 8 and Table 2, forward-fill EKF yielded a lower prediction error than the baseline error for all participants except for participant P8. For participants whose baseline prediction error was already below 10% (i.e., participants P3, P7, P9, and P10), the gains were understandably minor (see Figure 8). For participants P1, P2, and P4, forward-fill EKF led to significant improvements for mental state prediction error compared to estimations conducted without a state estimator, with decreases in prediction error of 4.0%, 13.1%, and 7.6%, respectively.

While the forward-fill EKF outperformed the forward-fill-evolving EKF for some participants, the opposite occurred for others. This variation suggests that the performance of the forward-fill-evolving EKF may be enhanced by personalizing parameter α per participant. Table 3 presents the average cumulative prediction error per participant using the forward-fill-evolving EKF under different values for α. The results show that the value of α that results in optimal performance of the forward-fill-evolving EKF varies across participants, further supporting the need for personalizing α.

B. COMPUTATIONAL EFFICIENCY

In addition to assessing the estimation accuracy, we also evaluated the computational efficiency of each state estimator. The average running times for one iteration of each state estimator were approximately 1.6 ms, 1.7 ms, and 1.9 ms, for the intermittent EKF, forward-fill EKF, and forward-fill-evolving EKF, respectively. Additionally, the maximum time for one iteration was 3.2 ms, 4.1 ms, and 5.1 ms for the intermittent EKF, forward-fill EKF, and forward-fill-evolving EKF, respectively. Such low computation times indicate that all filters are suitable for real-time applications in HSRI settings, where sampling intervals are typically on the order of several seconds [20].

C. CONVERGENCE ANALYSIS

As discussed in Section II-B6, general formal convergence guarantees are not tractable in our setting due to the nature of HSRI settings, which involve nonlinearity and depend on both context and user-specific personalization. In practice, empirical assessment of convergence behavior is more practical and informative in HSRIs. Therefore, we analyze the evolution of the trace of the covariance matrix $P(k)$ over time to evaluate whether the filters maintain bounded uncertainty during deployment. For the case of intermittent EKF, in steps where measurement is not available, the correction phase is not performed, and the actual covariance matrix $P(k)$ is

TABLE 2. Relative prediction error of the mental states for each participant per state estimator (the EKF variants developed in this paper, whose technical design is described in Section II-B) vs. no state estimator (introduced in [20]).

State estimator	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
No state estimator [20]	11.449	16.689	9.147	20.660	12.542	14.798	4.621	16.973	5.109	7.569
Intermittent EKF	11.443	16.814	9.127	20.714	12.545	14.833	4.608	16.987	5.149	7.548
Forward-fill EKF	7.472	3.620	8.604	13.047	12.455	12.489	4.231	17.888	4.469	6.554
Forward-fill-evolving EKF	7.813	4.899	8.369	14.396	12.411	12.459	4.290	17.564	4.521	6.496

TABLE 3. Impact of varying parameter α on the prediction error of the mental states for each participant when using the forward-fill-evolving EKF. The value of α that minimizes the overall prediction error for each participant is shown in bold.

α	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1[1]	**7.472**	**3.620**	8.604	**13.047**	12.455	12.489	4.231	17.888	**4.469**	6.554
1.1	7.482	3.689	8.592	13.093	12.453	12.489	4.234	17.878	4.474	6.552
2	7.813	4.899	8.369	14.396	12.411	**12.459**	4.290	17.564	4.521	6.496
5	9.742	8.925	**8.120**	17.838	12.011	13.096	4.449	17.190	4.644	5.965
10	10.541	12.138	8.325	19.143	**11.658**	14.804	4.401	17.415	4.772	**5.765**
100	11.173	16.035	8.89	20.449	11.941	16.972	**4.167**	**17.054**	5.069	5.911

[1] When $\alpha = 1$, the forward-fill-evolving EKF reduces to the forward-fill EKF.

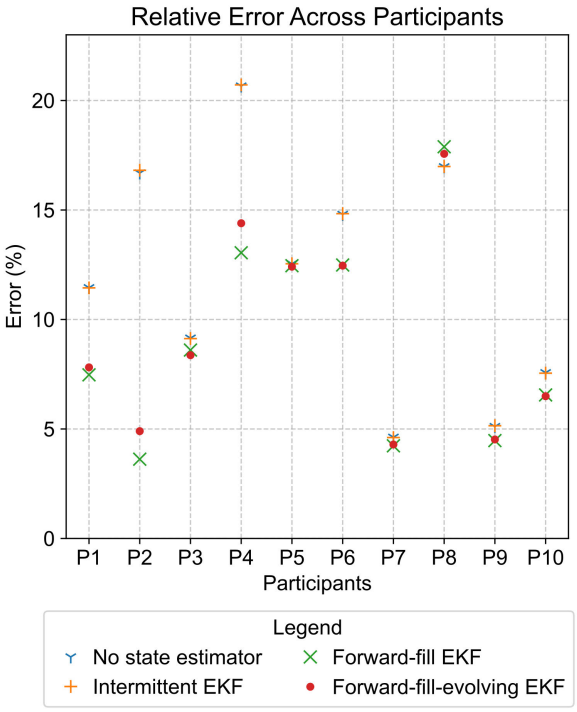


FIGURE 8. Relative prediction error of the mental states for each participant for each state estimator, compared to the case where no state estimator is deployed, introduced in [20]. The intermittent EKF, forward-fill EKF, and forward-fill-evolving EKF are adapted versions of EKFs [27] developed in this paper and described in Section II-B. Participants are represented by the letter "P" and a number on the horizontal axis.

not available (see (10c)). For these time steps, we report the predicted covariance matrix $P^-(k)$ (see (9b)) instead. The results are presented in Figure 9.

The covariance matrix of the intermittent EKF does not seem to converge for most participants, as its trace fluctuates over time. Between consecutive measurements, the trace of the predicted covariance matrix $P^-(k)$ increases: since the correction phase is not performed, process noise accumulates, and the predicted covariance matrix increases. Then, the trace of $P(k)$ only decreases when the correction phase is performed. This fluctuating behavior prevents the filter from reaching a stable error covariance and indicates that the filter does not exhibit practical convergence in the absence of frequent updates.

In contrast, the trace of the covariance matrix $P(k)$ for forward-fill EKF shows convergence for all participants. After an initial phase where the trace of $P(k)$ increases — likely due to a low initialization of $P(k)$ — it then decreases within the first 10 time steps and remains stable over time for all participants. This indicates that the filter maintains bounded uncertainty and exhibits converging behavior, making it suitable for HSRIs.

Finally, the forward-fill-evolving EKF exhibits better convergence behavior than intermittent EKF, but not as clear as forward-fill EKF. Despite applying both prediction and correction phases in each time step, forward-fill-evolving EKF still shows small fluctuations in $P(k)$ between measurements, likely due to the increase in the measurement noise covariance matrix $R(k)$ when measurements are unavailable. Given that a bounded $R(k)$ is often required for convergence of KFs, prolonged intervals without measurements can inflate uncertainty and reduce stability of forward-fill-evolving EKF. This reflects a trade-off between adaptivity and convergence in the estimation process that should be balanced in real-world applications.

Overall, these results suggest that forward-fill EKF is the most stable and practically suitable estimator for use in HSRIs, while intermittent EKF does not offer reliable convergence, and forward-fill-evolving EKF presents a trade-off between adaptivity and convergence. However, across all variants, the behavior of state estimators may vary depending on measurement sparsity, user-specific dynamics, and task conditions. As such, real-time heuristics that monitor the trace of the covariance matrix can serve as

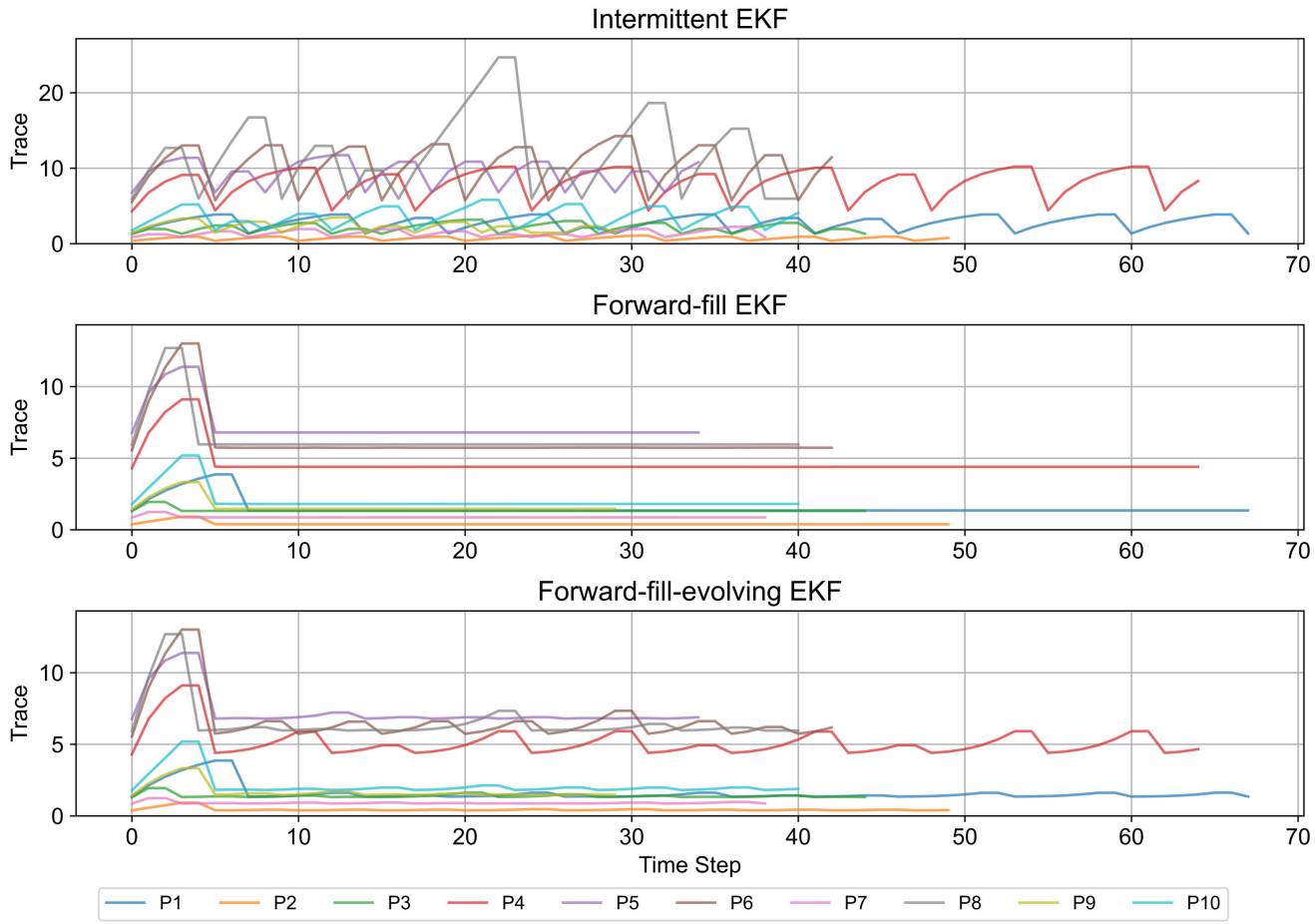


FIGURE 9. Trace of the state covariance matrix over time for 10 participants under three EKF variants: (a) intermittent EKF, (b) forward-fill EKF, and (c) forward-fill-evolving EKF. Each curve represents the trace of $P(k)$, for an individual participant — or the trace of $P^{-}(k)$ for intermittent EKF, when no correction is applied. A bounded or decreasing trace indicates stable estimation and practical convergence: (a) the intermittent EKF shows a sawtooth pattern and fails to converge without frequent updates; (b) the forward-fill EKF stabilizes rapidly after an initial transient and maintains bounded uncertainty; (c) the forward-fill-evolving EKF exhibits intermediate behavior, balancing adaptivity of noise weighting with convergence stability.

an important safeguard: they can provide early indicators of estimator degradation and enable adaptive responses, such as re-initialization or increased measurement frequency. Incorporating such heuristics will be critical for deploying model-based estimators in real-world HSRIs.

In summary, the forward-fill EKF was the best-performing state estimation method on the current dataset. It significantly reduced the prediction error of the mental states for most participants whose prediction error exceeded 10% without deploying a state estimator, and it showed converging behavior. While the forward-fill-evolving EKF achieved comparable performance, we discussed several factors that may explain why it did not outperform the forward-fill EKF in this case study. Furthermore, its convergence was slightly less stable than that of the forward-fill EKF and may degrade if measurements are significantly more infrequent than in the present study. Since both state estimators are simple to implement and improve the estimation of mental states of users, we recommend their use in HSRIs when measurements are infrequent, ideally in combination with heuristics that monitor filter reliability and adapt filter behavior accordingly. Specifically, when personalization of α is not feasible or measurements are particularly infrequent, the forward-fill EKF is preferable. Otherwise, the forward-fill-evolving EKF may provide improved performance.

Finally, while the results support the effectiveness of the proposed state estimators, we acknowledge that the number of participants is limited. This reflects a common constraint in HSRI experiments that involve multi-session protocols and personalized modeling. Additionally, since the evaluation was conducted offline, the performance of the estimators during live interactions remains to be validated. These factors should be considered when interpreting the generalizability of the findings.

V. CONCLUSION AND TOPICS FOR FUTURE RESEARCH

Just as humans maintain an awareness of each other's mental states during social interactions, SRs also require controllers that model the mental states of their users. In [20], MMM [24] — a dynamic mathematical model of human

perception, cognition, and decision-making — was integrated into a model-based controller to steer the behavior of an SR, enabling the robot to optimize the mental states of its users over the course of their interactions.

While the results of the proposed approach showed high promise for advancing social robotics, the prediction errors in estimating the mental states of some participants highlighted a key area for improvement. Given that the effectiveness of a model-based controller relies on having an accurate estimation of the current states of the system [43], reducing the state estimation and prediction errors is crucial for enhancing the overall performance of the controller.

Leveraging the existing model MMM, which mathematically describes the dynamic evolution of the relevant mental states, we proposed incorporating a model-based state estimator into the control loop of SRs to improve the accuracy of estimating and predicting these mental states. To this end, we developed and evaluated three versions of EKFs, tailored to estimate mental states of users in the context of human-robot social interactions.

We detailed how the parameters of these mental state estimators were derived and personalized to each participant, using data gathered in initial interactions with each participant. The proposed EKFs based on MMM were then evaluated on a separate portion of the same dataset [35] collected in experiments we previously conducted and reported in [20].

Since the evaluation of EKFs relied on previously collected data, our assessment was limited to quantifying improvements in the prediction error of mental states using the proposed EKFs. Although [20] demonstrated the potential of model-based control with MMM, it remains essential to investigate how improved state estimations actually affect the performance of the model-based controller and the overall quality of real-time interactions between humans and SRs.

As future work, we recommend implementing the proposed EKFs within the control loop during actual interaction sessions. This will allow assessing how the improved prediction of mental states impacts the behavior of SRs, as well as the interaction experience of their users. Additionally, a larger participant pool will be necessary to assess the generalizability of the proposed state estimators. Furthermore, we recommend integrating simple heuristics that can monitor convergence (e.g, establishing a threshold for the trace of the covariance matrix), and trigger adaptive responses such as re-initialization or increasing measurement frequency, if possible. Further investigation into the personalization of parameter α of the forward-fill-evolving EKF is also recommended, as our results suggest that tailoring this parameter per participant may significantly improve the performance of this state estimator.

While this work presents the first application of a model-based state estimator for invisible cognitive states in HSRIs, future research should explore alternative estimation frameworks — such as unscented Kalman filters, particle filters, or adaptive Kalman filters — within this context. Although applying these techniques will require adaptation to accommodate sparse and subjective measurements, doing so will potentially enable more comprehensive benchmarking and provide deeper insights into the performance of state estimators in cognitively driven interactive systems.

REFERENCES

[1] G. Gordon, S. Spaulding, J. Kory Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2016, vol. 30, no. 1, pp. 3951–3957.

[2] N. I. Arshad, A. S. Hashim, M. M. Ariffin, N. M. Aszemi, H. M. Low, and A. A. Norman, "Robots as assistive technology tools to enhance cognitive abilities and foster valuable learning experiences among young children with autism spectrum disorder," *IEEE Access*, vol. 8, pp. 116279–116291, 2020.

[3] B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola, and F. Shic, "Improving social skills in children with ASD using a long-term, in-home social robot," *Sci. Robot.*, vol. 3, no. 21, pp. 1–9, Aug. 2018.

[4] C. Clabaugh, K. Mahajan, S. Jain, R. Pakkar, D. Becerra, Z. Shi, E. Deng, R. Lee, G. Ragusa, and M. Matarić, "Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders," *Frontiers Robot. AI*, vol. 6, pp. 1–18, Nov. 2019.

[5] T. Ascensão and A. Jamshidnejad, "Autonomous socially assistive drones performing personalized dance movement therapy: An adaptive fuzzy-logic-based control approach for interaction with humans," *IEEE Access*, vol. 10, pp. 15746–15770, 2022.

[6] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, and S. Bermúdez I Badia, "Design, development, and evaluation of an interactive personalized social robot to monitor and coach post-stroke rehabilitation exercises," *User Model. User-Adapted Interact.*, vol. 33, no. 2, pp. 545–569, Apr. 2023.

[7] R. Feingold Polak and S. L. Tzedek, "Social robot for rehabilitation: Expert clinicians and post-stroke patients' evaluation following a long-term intervention," in *Proc. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, New York, NY, USA, Mar. 2020, pp. 151–160.

[8] S. Rossi, M. Larafa, and M. Ruocco, "Emotional and behavioural distraction by a social robot for children anxiety reduction during vaccination," *Int. J. Social Robot.*, vol. 12, no. 3, pp. 765–777, Jul. 2020.

[9] M. Luperto, J. Monroy, F.-A. Moreno, F. Lunardini, J. Renoux, A. Krpic, C. Galindo, S. Ferrante, N. Basilico, J. Gonzalez-Jimenez, and N. A. Borghese, "Seeking at-home long-term autonomy of assistive mobile robots through the integration with an IoT-based monitoring system," *Robot. Auto. Syst.*, vol. 161, Mar. 2023, Art. no. 104346.

[10] M. Schrum, C. H. Park, and A. Howard, "Humanoid therapy robot for encouraging exercise in dementia patients," in *Proc. 14th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2019, pp. 564–565.

[11] A. Kubota, E. I. C. Peterson, V. Rajendren, H. Kress-Gazit, and L. D. Riek, "JESSIE: Synthesizing social robot behaviors for personalized neurorehabilitation and beyond," in *Proc. 15th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, New York, NY, USA, Mar. 2020, pp. 121–130.

[12] I. Giorgi, F. A. Tirotto, O. Hagen, F. Aider, M. Gianni, M. Palomino, and G. L. Masala, "Friendly but faulty: A pilot study on the perceived trust of older adults in a social robot," *IEEE Access*, vol. 10, pp. 92084–92096, 2022.

[13] A. Tapus, M. Maja, and B. Scassellatti, "The grand challenges in socially assistive robotics," *IEEE Robot. Autom. Mag.*, vol. 14, no. 1, pp. 35–42, Jan. 2007.

[14] N. Céspedes, D. Raigoso, M. Múnera, and C. A. Cifuentes, "Long-term social human–robot interaction for neurorehabilitation: Robots as a tool to support gait therapy in the pandemic," *Frontiers Neurorobotics*, vol. 15, pp. 1–12, Feb. 2021.

[15] K. Shiarlis, J. Messias, and S. Whiteson, "Acquiring social interaction behaviours for telepresence robots via deep learning from demonstration," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, Mar. 2017, pp. 37–42.

[16] W.-R. Ko, J. Lee, M. Jang, and J. Kim, "End-to-end learning of social behaviors for humanoid robots," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 1200–1205.

[17] M. Maroto-Gómez, M. Malfaz, Á. Castro-González, S. Á. Arias, and M. Á. Salichs, "Deep reinforcement learning for the biologically inspired social behaviour of autonomous robots acting in dynamic environments," *IEEE Access*, vol. 12, pp. 180146–180160, 2024.
[18] Y. Gao, F. Yang, M. Frisk, D. Hernandez, C. Peters, and G. Castellano, "Social behavior learning with realistic reward shaping," 2018, *arXiv:1810.06979*.
[19] E. Bagheri, O. Roesler, H.-L. Cao, and B. Vanderborght, "A reinforcement learning based cognitive empathy framework for social robots," *Int. J. Social Robot.*, vol. 13, no. 5, pp. 1079–1093, Aug. 2021.
[20] M. Morão Patrício and A. Jamshidnejad, "Leveraging systems and control theory for social robotics: A model-based behavioral control approach to human–robot interaction," 2025, *arXiv:2504.21548*.
[21] D. Dell'Anna and A. Jamshidnejad, "SONAR: An adaptive control architecture for social norm aware robots," *Int. J. Social Robot.*, vol. 16, pp. 1–32, Oct. 2024.
[22] M. J. Matarić and B. Scassellati, "Socially assistive robotics," in *Springer Handbook of Robotics*, 1st ed., Heidelberg, Germany: Springer, 2016, pp. 1973–1993.
[23] Y. Tahir, J. Dauwels, D. Thalmann, and N. M. Thalmann, "A user study of a humanoid robot as a social mediator for two-person conversations," *Int. J. Social Robot.*, vol. 12, no. 5, pp. 1031–1044, Nov. 2020.
[24] M. L. M. Patrício and A. Jamshidnejad, "Dynamic mathematical models of theory of mind for socially assistive robots," *IEEE Access*, vol. 11, pp. 103956–103975, 2023.
[25] G. Welch and G. Bishop, "An introduction to the Kalman filter," Univ. North Carolina, Chapel Hill, NC, USA, Tech. Rep. 95-041, 1995.
[26] Y. Kim and H. Bang, "Introduction to Kalman filter and its applications," in *Introduction and Implementations of the Kalman Filter*, 1st ed., Rijeka, Croatia: IntechOpen, 2018, ch. 2.
[27] D. Simon, *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*, 1st ed., Hoboken, NJ, USA: Wiley, 2006.
[28] M. S. Grewal, *Kalman Filtering: Theory and Practice With MATLAB*, 4th ed., Hoboken, NJ, USA: Wiley, 2015.
[29] M. Nakamura, "Relationship between steady state Kalman filter gain and noise variances," *Int. J. Syst. Sci.*, vol. 13, no. 10, pp. 1153–1163, Oct. 1982.
[30] F. Wu, H. Luo, H. Jia, F. Zhao, Y. Xiao, and X. Gao, "Predicting the noise covariance with a multitask learning model for Kalman filter-based GNSS/INS integrated navigation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.
[31] R. Mehra, "On the identification of variances and adaptive Kalman filtering," *IEEE Trans. Autom. Control*, vol. AC-15, no. 2, pp. 175–184, Apr. 1970.
[32] P. Matisko and V. Havlena, "Noise covariance estimation for Kalman filter tuning using Bayesian approach and Monte Carlo," *Int. J. Adapt. Control Signal Process.*, vol. 27, no. 11, pp. 957–973, Nov. 2013.
[33] B. Feng, M. Fu, H. Ma, Y. Xia, and B. Wang, "Kalman filter with recursive covariance estimation—Sequentially estimating process noise covariance," *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6253–6263, Nov. 2014.
[34] B. J. Odelson, M. R. Rajamani, and J. B. Rawlings, "A new autocovariance least-squares method for estimating noise covariances," *Automatica*, vol. 42, no. 2, pp. 303–308, Feb. 2006.
[35] M. L. M. Patrício and A. Jamshidnejad, "Data from experiments of 'leveraging systems and control theory for social robotics: A model-based behavioral control approach to human–robot interaction," 4TU.ResearchData, Delft Univ. Technol., Delft, The Netherlands, Tech. Rep., Mar. 2025, doi: 10.4121/ccadc914-9502-46d6-9ba5-fef581f2933f.v1.
[36] M. Boutayeb, H. Rafaralahy, and M. Darouach, "Convergence analysis of the extended Kalman filter used as an observer for nonlinear deterministic discrete-time systems," *IEEE Trans. Autom. Control*, vol. 42, no. 4, pp. 581–586, Apr. 1997.
[37] S. Bonnabel and J.-J. Slotine, "A contraction theory-based analysis of the stability of the deterministic extended Kalman filter," *IEEE Trans. Autom. Control*, vol. 60, no. 2, pp. 565–569, Feb. 2015.
[38] K. Sueki, S. Nishizawa, T. Yamaura, and H. Tomita, "Precision and convergence speed of the ensemble Kalman filter-based parameter estimation: Setting parameter uncertainty for reliable and efficient estimation," *Prog. Earth Planet. Sci.*, vol. 9, no. 1, p. 47, Sep. 2022.
[39] SoftBank Robotics. *NAO Robot*. Accessed: Apr. 14, 2025. [Online]. Available: https://www.softbankrobotics.com/emea/en/nao
[40] A. Robaczewski, J. Bouchard, K. Bouchard, and S. Gaboury, "Socially assistive robots: The specific case of the NAO," *Int. J. Social Robot.*, vol. 13, no. 4, pp. 795–831, Jul. 2021.
[41] T. Alhmiedat and M. Alotaibi, "Design and evaluation of a personal robot playing a self-management for children with obesity," *Electronics*, vol. 11, no. 23, p. 4000, Dec. 2022.
[42] D. Pandey, A. Subedi, and D. Mishra, "Improving language skills and encouraging reading habits in primary education: A pilot study using NAO robot," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2022, pp. 827–832.
[43] M. Morari and J. H. Lee, "Model predictive control: Past, present and future," *Comput. Chem. Eng.*, vol. 23, pp. 667–682, Mar. 1999.

MARIA L. MORÃO PATRÍCIO received the B.Sc. degree in aerospace engineering from the Instituto Superior Técnico, Portugal, in 2019, and the M.Sc. degree (cum laude) in control and operations from the Aerospace Engineering Faculty, Delft University of Technology, The Netherlands, in 2021, where she is currently pursuing the Ph.D. degree. Her research focuses on the integration of explicit human models into control frameworks, aiming to enable social robots to interact intelligently and adaptively with humans. She is particularly interested in model-based control, cognitive architectures, and the design of human-aware autonomous systems.

ANAHITA JAMSHIDNEJAD received the Ph.D. degree (cum laude) from Delft University of Technology (TU Delft), The Netherlands, in 2017. She is currently Assistant Professor with TU Delft, leading the Mathematical Decision Making Group and directing the AI*MAN Laboratory. Her main research interests include systems theory for modeling human cognition, model-based predictive control frameworks, fuzzy logic, and integrated/hierarchical control paradigms, applied to autonomous and social robots.