Expert Systems With Applications 323 (2026) 132449

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Interpretable modeling of human decision-making from user interactions in a dynamic stabilization task

Ildikó Horváth [a,b,*], Anna Sudár [a,b,c], Ádám B. Csapó [a,b,c]

[a] Corvinus Institute for Advanced Studies, Corvinus University of Budapest, Fovam square 8, Budapest, 1093, Hungary
[b] Institute of Data Analytics and Information Systems, Corvinus University of Budapest, Fovam square 8, Budapest, 1093, Hungary
[c] Hungarian Research Network, Piarista u. 4, Budapest, 1052, Hungary

ARTICLE INFO

Keywords:
Fuzzy linguistic modeling
Decision-making
Decision asymmetries
Cognitive biases

ABSTRACT

This paper introduces the Extended Takagi-Sugeno Fuzzy Model Transformation (ETSFM), a non-parametric, data-driven framework grounded in the TS fuzzy model transformation combined with Close-to-Normal (CNO) and Inverted Relaxed Normal (IRNO) transformations. Without requiring pre-defined membership templates, ETSFM turns interaction data into compact, interpretable fuzzy linguistic rule sets capable of explaining human decision strategies in black-box and gray-box dynamic environments. Applied to an abstract stabilization task as a demonstrative example, the approach isolates three robust patterns in decision-making: reliance on *self-stabilization regions* (state-space zones where small or no inputs are sufficient to maintain control), *transition zones* where control effort steeply increases, and *blind spots* characterized by repeated failures. The systematic asymmetry uncovered in human control manifests as a directional bias: kinematically mirrored states between left and right elicit qualitatively different control decisions. The resulting rule sets are low-complexity yet high-fidelity, capturing the dominant decision logic while remaining auditable. Based on these results, the paper provides two contributions to the literature. Methodologically, we provide a principled workflow for transforming raw data into transparent, linguistically labeled rules without relying on pre-defined fuzzy variables or fuzzy membership templates. In terms of human decision behavior in black-box and gray-box environments, we uncover a systematic asymmetry in human control that endures across a range of dynamic conditions. In a practical sense, the framework can serve as a diagnostic tool for benchmarking human-AI decision systems: it is capable of mapping error-prone states, supporting targeted training and UI refinement, while permitting side-by-side comparison with an optimized controller. The method is transfer-ready to domains where decisions are inferred from interaction data under partial model knowledge, such as driver assistance or process control.

1. Introduction

Human decision-making in dynamic control tasks is a rich but underexplored area at the intersection of cognitive science, control theory, and human-computer interaction. When the underlying system dynamics are hidden from the user – as in many real-world black-box and gray-box environments – decision-making can no longer rely on explicit mechanical knowledge. Instead, it must draw on internal cognitive representations built from sparse feedback and prior experience (Johnson-Laird, 1983; Norman, 2014). Understanding the structure of these representations, and how they lead to systematic control patterns and errors, has clear implications for the design of decision-support systems, adaptive interfaces, and human-AI collaboration frameworks.

In this paper, our goal is to develop a methodology for translating measurement data – specifically, data obtained from users interacting with black-box and gray-box dynamic environments – into interpretable fuzzy linguistic rules. The core research question is: *How can raw user interaction data from such settings be systematically transformed into compact, auditable fuzzy rule sets that reveal the structure of human decision-making, without relying on pre-defined fuzzy variables or membership function templates?* Existing approaches to data-driven fuzzy rule generation either require pre-defined membership shapes (e.g., Gaussian or triangular) or resort to clustering heuristics, neither of which readily extends to cognitive decision data (Angelov & Yager, 2011; Lughofer, 2008). The present work addresses this gap directly.

We apply the developed methodology, referred to as the **Extended TS Fuzzy Model Transformation for Rule Set Generation** (hereafter abbreviated as the "*ETSFM transformation*", or simply "*ETSFM*"), to the topic of understanding how humans make decisions when trying to balance a dynamic system whose underlying dynamics is unknown to them.

* Corresponding author.
E-mail addresses: ildiko.horvath@uni-corvinus.hu (I. Horváth), anna.sudar@uni-corvinus.hu (A. Sudár), adambalazs.csapo@uni-corvinus.hu (Á.B. Csapó).

https://doi.org/10.1016/j.eswa.2026.132449
Received 26 January 2026; Received in revised form 10 April 2026; Accepted 10 April 2026
Available online 14 April 2026

ETSFM is built upon the Tensor Product (TP) / Takagi-Sugeno fuzzy model transformation framework (Baranyi, 2016, 2023a) and employs Close-to-Normal (CNO) and Inverted Relaxed Normal (IRNO) transformations to extract non-parametric, data-driven antecedent fuzzy sets from the raw data (Baranyi et al., 2017). The resulting rule sets take the form of linguistically labeled IF-THEN statements whose structure is determined entirely by the data, without pre-conceived membership function templates.

Our analysis reveals three distinct control regimes: (i) *self-stabilization regions* – central state-space zones in which small or zero control inputs are sufficient to maintain the system within the acceptable region; (ii) *transition zones*, where control effort steeply increases; and (iii) *blind spots*, characterized by repeated failures to identify a successful way of intervening. We also identify systematic asymmetries in a gray-box dynamic control task that reflect inherent imbalances in human control. We hypothesize that the proposed methodology and resulting concepts could be applied to a broader range of human-AI control problems, especially where limited theoretical understanding motivates a data-driven approach.

From a theoretical perspective, the proposed methodology advances the state of the art by providing a non-parametric model capable of transforming raw behavioral measurements into an explainable set of fuzzy IF-THEN rules – without requiring pre-determined templates regarding the support or shape of the underlying fuzzy membership functions.

From a practical perspective, the paper focuses on modeling human action-reaction behavior when stabilizing a second-order dynamic system that, unknown to users, exhibits characteristics similar to an inverted pendulum in its upright position. In this scenario, participants interact with an abstract visual simulation on a computer screen and respond by entering control inputs via a keyboard. Importantly, users do not receive direct visual-proprioceptive-motor feedback, nor do they engage fast cerebellar motor-control pathways. Instead, decision-making relies primarily on higher-order cognitive processes rather than instinctive motor responses, making the task particularly suitable for analyzing high-level human control strategies.

The key contributions of this paper are twofold. First, from a methodological standpoint, we introduce the ETSFM framework, a fully non-parametric, data-driven approach that extends the TS fuzzy model transformation to the domain of cognitive decision-making – an application area not previously addressed in either the control theory or cognitive modeling literature. A distinctive feature of the approach is the use of CNO / IRNO transformations to shape linguistically interpretable, orthogonal fuzzy rule regions without iterative supervised learning. Second, from an empirical standpoint, we provide the first systematic fuzzy rule based characterization of human control behavior in a gray-box second-order dynamic stabilization task, revealing self-stabilization zones, transition regions, blind spots and a persistent directional asymmetry across participants.

The paper is structured as follows. After reviewing the relevant literature (Section 2), the mathematical notations as well as key definitions are presented in Section 3. This lays the foundation for the newly proposed ETSFM transformation, which is introduced and analyzed – including rule set validation and computational complexity – in Section 4. Building on this methodological foundation, the application of the methodology in gray-box dynamic control assessment is described in Section 5, followed by the measurement procedure used in our particular case study in Section 6. Results and further analyses are provided in Section 7, and the discussions are provided in Section 8.

2. Literature review

In this section, an overview is provided on past research dealing with data-driven rule set generation – particularly in the context of human behavior – as well as with high-level models of human and AI decision making.

2.1. Data-driven fuzzy rule set generation

Several works in the past have explored the extraction of fuzzy rules directly from experimental or numerical data, aiming to bridge data-driven modeling and interpretable fuzzy inference.

Early efforts include a rough set theory based approach by Plonka and Mrozek (1995) to derive decision rules from human demonstrations of inverted-pendulum control. Zapata et al. (1999) modeled the human operator through an intermediary ARMA system to smooth experimental data before generating fuzzy rules. Both of these approaches resulted in fuzzy rule sets, albeit based on pre-defined hyperparameters or surrogate models. In a somewhat different approach, Wu and Chen (1999) presented an explicitly data-driven approach based on α-cuts of fuzzy equivalence relations, which automatically partitioned the input-output space into fuzzy regions.

Further advances came with clustering-based methods, which enabled geometric discovery of structure through clustering in data space. Chiu (1997) introduced subtractive clustering to identify high-density regions in the input-output space, each defining a fuzzy rule whose antecedents were modeled by Gaussian membership functions arrived at through optimization. Subsequent works extended this idea through adaptive and hybrid schemes. Zarandi et al. (2004) coupled Gustafson-Kessel fuzzy clustering with neural network parameter optimization and sequential quadratic programming, while Lughofer (2008) introduced incremental and evolving learning mechanisms through the FLEXFIS framework that continuously adapt both premise and consequent parameters. Makrehchi and Kamel (2011) proposed an alternative optimization approach based on information-theoretic criteria and genetic algorithms for shaping trapezoidal memberships. In all of these methods, clustering or optimization determines the structure of the rule base, but the representation of fuzzy sets remains parametric, typically assuming Gaussian, trapezoidal or triangular forms whose parameters must be tuned.

Angelov and Yager (2011) extended the data-driven paradigm by removing the need for explicit membership functions, by representing antecedents as data clouds in their ANYA framework. This generalizes the clustering logic introduced by Chiu (1997), moving from parametric representations of clusters to purely density-based antecedents. Rule activation is computed as a function of local and global data density, allowing the rule base to evolve adaptively.

More recently, deep learning approaches have been applied to fuzzy rule generation. Methods based on neuro-fuzzy networks (Jang, 1993) and deep rule-based networks combine the representational power of neural networks with the linguistic interpretability of fuzzy systems (Gu & Angelov, 2020; Zhang et al., 2025). While these approaches can achieve high accuracy, they typically learn fuzzy sets through gradient-based optimization over pre-defined membership function templates, and the resulting rule bases may lose interpretability as the number of rules grows (Pickering et al., 2025). Furthermore, they require labeled training data and substantial parameter tuning, making them less suited to exploratory behavioral analysis where the goal is structure discovery rather than predictive optimization.

The approach proposed here diverges fundamentally from prior solutions based on clustering or density estimation, or neural-network optimization. Rather than identifying clusters, estimating data densities, or fitting parameterized membership functions through gradient descent, the Tensor Product (TP) and Close-to-Normal (CNO) transformations extract fuzzy rules directly through matrix or tensor decomposition. The method is fully nonparametric and analytic, requiring no pre-defined membership functions, clustering hyperparameters or surrogate models. The CNO transformation ensures that each rule corresponds to a unique, orthogonal region of the state space, providing interpretable partitions that naturally highlight dominant decision regions and expose ambiguity without any iterative optimization. Combinatorial rule generation is then executed over only those regions in the input space that are structurally identified through peaks in the previously identi-

fied membership functions. Unlike parametric approaches (which require strong prior assumptions about membership shape), or neural-network approaches (which sacrifice transparency for expressiveness), the TP/CNO based method provides a transparent path from raw data to linguistic rules with minimal assumptions, at the cost of requiring grid-structured input-output data.

2.2. Human and AI decision making

Prior research points to significant differences in human and artificial intelligence (AI) based decision making. Human decision-makers often rely on cognitive flexibility, common sense, and contextual knowledge, while AI-based controllers excel at statistical pattern recognition, optimization, and efficient handling of large data sets (Rastogi et al., 2023). This underscores the difficulty of designing AI controllers capable of achieving human-like adaptability while maintaining optimal control performance.

Classical research in cognitive science and behavioral economics has documented systematic biases in human decision-making under uncertainty and in dynamic environments. Prospect theory (Kahneman & Tversky, 1979) establishes that people evaluate outcomes relative to a reference point and are asymmetrically sensitive to gains and losses, a finding relevant to understanding why human decision-making systematically diverges from mathematically optimal control strategies. More broadly, research on naturalistic decision-making (Klein, 1999) and the dual-process model of System I versus System II thinking (Kahneman, 2011) suggests that high-level cognitive decisions in complex, time-pressured environments are often governed by heuristics rather than deliberate optimization. These findings motivate the use of interpretable rule-based representations, such as the fuzzy rule sets generated by ETSFM, for capturing the structure of human decision logic.

Studies on dynamic decision-making (DDM) emphasize that humans often employ instance-based learning, which means that decisions are made based on past experiences stored in memory (Gonzalez, 2022). These cognitive mechanisms contrast with AI-driven models, which operate based on explicit mathematical formulations and optimization techniques. Kloosterman et al. (2020) found that humans dynamically adjust their decision biases based on contextual signals, suggesting that adaptability plays a crucial role in human decision-making. However, it is challenging to design an appropriate measurement environment for assessing dynamic adaptability. The environment must be sufficiently complex and variable to ensure that learning is neither too simple nor overly predictable. At the same time, the user interface (UI) should remain intuitive to prevent additional cognitive load that distracts from problem-solving.

Our research focuses on black-box and gray-box scenarios where the underlying dynamic system is at least partially opaque to users. In this context, an example experimental design is developed based on an inverted pendulum system. Participants perform a seemingly simple task: controlling the movement of a ball within predefined limits. However, the underlying control dynamics remain hidden, making the stabilization process more complex than it initially appears. This mirrors real-world scenarios in which decision-makers operate without explicit knowledge of system rules.

The existing literature suggests that in such scenarios, people exhibit decision biases that vary with cognitive load and uncertainty levels (Kloosterman et al., 2020). EEG-based studies have shown that neural variability also correlates with shifts in decision bias, with higher neural entropy in frontal regions leading to more flexible decision-making strategies (Kloosterman et al., 2020).

Also, past research has shown that humans and AI exhibit different decision biases and processing speeds in dynamic environments (Ryu et al., 2021). While deep learning-based decision models can predict future states faster than humans, human decision-makers often outperform AI when faced with sudden environmental changes, as they are able to adapt heuristics in ways that AI has difficulty replicating (Ryu et al., 2021).

By contrasting human-derived fuzzy rule sets with AI-based controllers, we expect to be able to quantify logical differences in decision-making and explore approaches to integrating human-style reasoning into automated decision-support systems. A key challenge in this context is the development of decision-making models that combine human intuitive adaptability with the efficiency and accuracy of machine-based methods. Fuzzy logic has shown promise in bridging this gap, offering interpretability and the ability to handle uncertain or partially defined environments where intuitive rather than strictly rule-based decision-making prevails (Baranyi, 2004, 2020b, 2023b, 2024; Zadeh, 1996)

Active synthesis, which enables the dynamic simulation and replication of decision-making processes, is also increasingly vital for decision-support systems. While agent-based modeling (Epstein, 2012) and reinforcement learning (Sutton & Barto, 1998) have explored active simulation, the explicit integration of fuzzy logic for interpretability and adaptive simulation presents novel opportunities. In particular, prioritizing fuzzy rules and structuring the decision-making mechanisms in a transparent manner can significantly enhance comprehensibility and traceability (Guillaume, 2001).

3. Mathematical notations and definitions

Before introducing the proposed Extended TS Fuzzy Model Transformation for Rule Set Generation (ETSFM transformation), in this section we define the mathematical notations used in the paper, as well as the TS fuzzy model based on which the ETSFM transformation is developed.

3.1. Notations

The following notations are used in the paper:

- Indices are denoted by lowercase letters of the alphabet, e.g. $i, j, k, l, n \ldots$ with corresponding upper bounds denoted using their capitalized version, e.g. $I, J, K, L, N \ldots$.
- Scalars, vectors, matrices and tensors are denoted as $s \in \mathbb{R}$, $\mathbf{s} \in \mathbb{R}^J$, $\mathbf{S} \in \mathbb{R}^{J_1 \times J_2}$, $\mathcal{S} \in \mathbb{R}^{J^N}$. In the case of tensors, the notation $\mathcal{S} \in \mathbb{R}^{I^N \times J^2}$ is equivalent to $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N \times J_1 \times J_2}$.
- $[\mathcal{S}]_{index}$ addresses elements in $\mathcal{S}$, e.g. $[\mathcal{S}]_{i_1,i_2,\ldots i_N} = \mathbf{S}_{i_1,i_2,\ldots i_N} \in \mathbb{R}^{J^2}$ of $\mathcal{S} \in \mathbb{R}^{I^N \times J^2}$;
- $[\mathcal{S}]_{i_1/i_2/\ldots}$ is a vector containing the indexed elements as $\begin{bmatrix}[\mathcal{S}]_{i_1} & [\mathcal{S}]_{i_2} & \ldots\end{bmatrix}$;
- $w_i(p)$ denotes Ruspini-partitioned membership functions such that $\forall p : \sum_{i=1}^{I} w_i(p) = 1$ and $\forall i, p : 0 \le w_i(p)$.

Definition 1 (TS Fuzzy Model). *Consider a set of fuzzy rules in the form as in Eq. (1):*

$$IF\ A_{1,i_1}\ AND\ A_{2,i_2}\ THEN\ \mathbf{S}_{i_1,i_2}. \tag{1}$$

where the antecedent fuzzy sets A_{n,i_n} are defined by Ruspini-partitioned membership functions $w_{n,i_n}(p_n)$. The consequents are linear time and parameter invariant system matrices $\mathbf{S}_{i_1,i_2} \in \mathbb{R}^{J^2}$ which are also referred to as vertex systems, consequent vertices or simply vertices. As shown in Eq. (2), if the observation fuzzy sets are singletons located at p_n and the product-sum-gravity inference operator is used, then the transfer function of the TS fuzzy model takes the form (Baranyi, 2016, 2023a; Baranyi et al., 2017; Yam et al., 1999):

$$\mathbf{S}(\mathbf{p}) = \sum_{i_1}^{I_1} \sum_{i_2}^{I_2} w_{1,i_1}(p_1) w_{2,i_2}(p_2) \mathbf{S}_{i_1,i_2}. \tag{2}$$

An important property of Eq. (2) is that, for any input, the output remains within the convex hull defined by the vertices, i.e. $\mathbf{S}(\mathbf{p}) \in co\{\mathbf{S}_{i_1,i_2}\}$, which is guaranteed by the Ruspini-partitioning. Functions $w_{1,i}(p_1)$ and $w_{2,i}(p_2)$ are referred to as the antecedent system. Matrices $\mathbf{S}_{i_1,i_2}$ are referred to as the consequent or the vertex system.

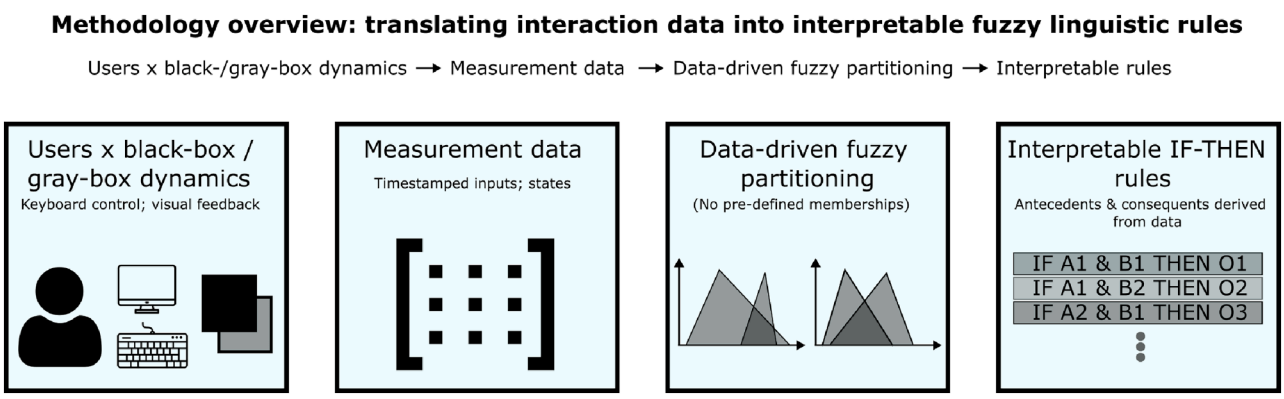

Fig. 1. Methodological overview – from interaction data to interpretable IF-THEN rules.

The foundational literature of the original TS fuzzy model transformation can be found in Baranyi (2004, 2014, 2016, 2020a,b, 2023a,b, 2024), Baranyi et al. (2017).

4. Extended TS fuzzy model transformation for rule set generation

This section introduces the extended version of the TS fuzzy model transformation that can be used for the extraction of linguistically interpretable rules from measurement data, a process referred to as fuzzy rule set generation. Fig. 1 provides a high-level illustration of the four-step workflow. The theoretical foundation of the transformation lies in the well-established TS fuzzy model transformation and TP model framework (Baranyi, 2016, 2023a; Baranyi et al., 2017), which this paper extends to the non-parametric extraction of fuzzy rules from behavioral data. The extension is non-trivial, and can be summarized in the following points:

- ETSFM removes the requirement that the input data come from an analytically defined system model, accepting instead raw empirical measurements arranged on a grid;
- ETSFM introduces a new output-clustering step (See Step 3 below) based on SVD of a restructured output matrix, enabling principled, parameter-free identification of consequent levels;
- ETSFM provides a complete workflow from behavioral data to auditable linguistic rules, which has not previously been applied in the context of cognitive decision-making research

Assume that we have a system with two inputs and one output. Based on the system, we generate a matrix $\mathbf{T}$ that represents a sampling of the system over a rectangular grid in input value combinations. Thus the element $[\mathbf{T}]_{m_1,m_2}$ is a measured value over input $[\mathbf{g}_1]_{m_1}$ on the first dimension and $[\mathbf{g}_2]_{m_2}$ on the second dimension. Here vectors $\mathbf{g}_1$ and $\mathbf{g}_2$ define the values of the grid on the input dimensions.

The following set of operations can be carried out in such 2-dimensional cases, as well as in higher-dimensional cases where $\mathcal{T}$ is a tensor indexed by more than 2 dimensions. For the sake of simplicity, we focus on the 2-dimensional case, with the general observation that in the higher-dimensional case the use of higher-order singular value decomposition (HOSVD) can be substituted for SVD, and methods like higher-order orthogonal iteration (HOOI) can be used instead of simply removing singular values for dimensionality reduction, as the latter approach would not guarantee minimal-error reconstruction in the higher-dimensional case (Ishteva et al., 2008). We further note that besides using HOSVD / HOOI instead of SVD / 2-dimensional rank reduction, all four steps outlined below remain unchanged in the higher-dimensional case.

- **Step 1: Principal component analysis**

  Rationale: The empirical data matrix $\mathbf{T}$ is typically a noisy, full-rank object. Singular value decomposition (SVD) (Beltrami, 1873) provides the best low-rank approximation in the least-squares sense (Eckart-Young theorem), separating the dominant structure from measurement noise. The resulting left and right singular vectors serve as the raw antecedent weighting functions before normalization. Retaining only the top I_1 singular values is the optimal rank-I_1 approximation, ensuring minimal reconstruction error.

  Execute SVD (Beltrami, 1873) on matrix T as shown in Eq. (3):

$$\mathbf{T} \overset{svd}{=} \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{3}$$

  The values of the resulting matrices $\mathbf{U}$ and $\mathbf{V}$ range between -1 and 1, with both $\mathbf{U}$ and $\mathbf{V}$ being orthonormal matrices. The diagonal matrix $\mathbf{D}$ contains the singular values σ_r in descending order, and the number of non-zero singular values corresponds to the rank R of matrix $\mathbf{T}$. Therefore, $\mathbf{T}$ can be expressed as shown in Eq. (4):

$$\mathbf{T} = \sigma_1[\mathbf{U}]_1[\mathbf{V}]_1^T + \sigma_2[\mathbf{U}]_2[\mathbf{V}]_2^T + \ldots + \sigma_R[\mathbf{U}]_R[\mathbf{V}]_R^T, \tag{4}$$

  where $[\mathbf{U}]_r$ and $[\mathbf{V}]_r$ are the r-th column vectors. In the subsequent steps, each $[\mathbf{U}]_r$ will be transformed into the membership function of an antecedent fuzzy set. Therefore, the number of antecedent fuzzy sets can be determined at this stage. If the goal is to retain only $I_1 \leq R$ antecedent fuzzy sets along e.g. the first dimension, the first I_1 singular values are kept as shown in Eq. (5):

$$\mathbf{T} \approx \sigma_1[\mathbf{U}]_1[\mathbf{V}]_1^T + \sigma_2[\mathbf{U}]_2[\mathbf{V}]_2^T + \ldots + \sigma_R[\mathbf{U}]_{I_1}[\mathbf{V}]_{I_1}^T, \tag{5}$$

  Since the singular values are arranged in descending order, discarding the smaller singular values along with their associated column vectors in $\mathbf{U}$ and $\mathbf{V}$ introduces the minimal possible error (at least in the 2-dimensional case). Let the final reduced matrices be denoted as $\mathbf{U}^r$ and $\mathbf{V}^r$.

- **Step 2: Shaping Linguistically Meaningful Antecedents**

  Rationale: The raw singular vectors from Step 1 can take negative values and do not sum to unity, so they cannot be interpreted as fuzzy membership functions. The CNO and IRNO transformations map them onto $[0, 1]$, while preserving the Ruspini partition property ($\sum_i w_i(p) = 1$). The CNO transformation additionally ensures that each resulting fuzzy set achieves a maximum value as close as possible to 1 at some point, making it possible to identify a "dominant" operating region for each rule. This is what enables the assignment of linguistic labels (e.g., "*small angle*", "*large velocity*") to the antecedents. The IRNO transformation serves the complementary role of identifying regions where a given rule is least dominant, exposing boundary and transition regions. Together, the two variants provide complementary views of the same decision landscape, enabling both a "*where does this rule fire*" and a "*where does this rule not fire*" type analysis.

  First we focus on the antecedent fuzzy sets on the inputs. Following the concept of the TS fuzzy model transformation, let the elements of the columns of $\mathbf{U}^r$ and $\mathbf{V}^r$ be considered as the values of

weighting functions over the grid, defined as shown in Eq. (6):

$$w_{1,i_1}([\mathbf{g}_1]_{m_1}) = [\mathbf{U}^r]_{m_1,i_1} \quad \text{and} \quad w_{2,i_2}([\mathbf{g}_2]_{m_2}) = [\mathbf{V}^r]_{m_2,i_2}. \tag{6}$$

Then the piecewise linear interpolation between the grid indices defines continuous weighting functions over all ranges of $[g_{1,m_1}, g_{1,m_{1+1}}]$ and $[g_{2,m_2}, g_{2,m_{2+1}}]$ as shown in Eqs. (7) and (8):

$$w_{1,i_1}(x_1) = (1-\lambda_1)[\mathbf{U}^r]_{m_1,i_1} + \lambda_1[\mathbf{U}^r]_{m_1+1,i_1} \tag{7}$$

$$w_{2,i_2}(x_2) = (1-\lambda_2)[\mathbf{V}^r]_{m_2,i_2} + \lambda_2[\mathbf{V}^r]_{m_2+1,i_2} \tag{8}$$

where λ_1 and λ_2 are defined as shown in Eq. (9):

$$\lambda_1 = \frac{x_1 - g_{1,m_1}}{g_{m_1+1} - g_{m_1}} \quad \text{and} \quad \lambda_2 = \frac{x_2 - g_{2,m_2}}{g_{2,m_2+1} - g_{2,m_2}}. \tag{9}$$

To ensure that the weighting functions can be interpreted as fuzzy sets, the values of $\mathbf{U}^r$ and $\mathbf{V}^r$ must be transformed onto the range of $[0, 1]$. Various transformation methods for this purpose are discussed in the literature related to the TS fuzzy model transformation (Baranyi et al., 2017). In the present approach, the CNO-type and the IRNO-type transformations are selected.

The CNO-type (close to normal) approach involves first transforming $\mathbf{U}^r$ into a matrix $\mathbf{U}^C$ (where the superscript C denotes CNO) such that its elements fall within the range $[0, 1]$, and the sum of the elements in each row equals one, as expressed in Eq. (10):

$$\mathbf{U}^C\mathbf{1} = \mathbf{1}. \tag{10}$$

A numerical iteration is then executed to adjust matrix $\mathbf{U}^C$, ensuring that the maximum value in as many columns as possible equals one, while the maximum value in all other columns (where this is not possible) is as close to one as possible.

The IRNO-type (inverted relaxed normal) transformation results in matrix $\mathbf{U}^I$ (where the superscript I denotes IRNO) whose elements are non-negative and satisfy Eq. (11):

$$\mathbf{U}^I\mathbf{1} = \mathbf{1} \tag{11}$$

holds as in the case of $\mathbf{U}^C$. However, the transformation identifies the minimum possible value such that all columns share this common minimum.

Applying the above piecewise linear interpolation technique to $\mathbf{U}^C$ and $\mathbf{U}^I$ yields weighting functions $\mathbf{w}_n^C(p_n)$ and $\mathbf{w}_n^I(p_n)$ that can be interpreted as the membership functions of the antecedent fuzzy sets. The maxima of functions $\mathbf{w}_n^C(p_n)$ determined by $\mathbf{U}^C$ reach or approach 1, effectively highlighting the dominant points or regions of the input covered by the fuzzy sets. This characteristic enables the meaningful association of linguistic labels to these elements, enhancing interpretability (see later in Section 7). In the case of $\mathbf{w}_n^I(p_n)$, the resulting membership functions of the antecedent fuzzy sets primarily indicate regions where the fuzzy sets are not dominant, serving an opposite role to that of the CNO-type fuzzy sets.

- **Step 3: Shaping linguistically meaningful consequents**

  Rationale: Having identified the antecedent structure, we need a principled way to group output values into a small number of interpretable consequent levels (e.g., "*small*", "*moderate*", "*large*" force). The matrix $\mathbf{T}^O$ restructures all measured output values in increasing order, appending their input coordinates so that SVD can detect dominant response levels in the output space. The row-wise sorting ensures that the dominant row-space structure identified by SVD corresponds to clusters of output levels, enabling the columns in U^O to be interpreted as weighting functions over the output range. SVD is used here – rather than k-means or Gaussian mixture models – because it provides an algebraically consistent, parameter-free decomposition of the output structure, aligned with the same mathematical framework used for the antecedents in Step 1. This consistency is important: it ensures that the number of consequent clusters can be chosen in the same rank-selection framework of the antecedents, and that the resulting consequent levels are derived from the global structure of the output distribution rather than from a locally optimal cluster assignment. The row-replication scheme for filling the matrix (repeating the last available input pair for rows with fewer matches) is a conservative choice that avoids introducing fictitious data points.

  Let us now turn our focus to the classification of the output values. To this end, we define matrix $\mathbf{T}^O$ (superscript O refers to the output) that is a restructured variant of $\mathbf{T}$, where the first column contains all the values of $\mathbf{T}$ in increasing order. Thus, the first element of each row is the measured output value. The second and third elements represent the input values along the x_1 and x_2 dimensions that correspond to the output value. If multiple pairs of x_1 and x_2 are present to the same output value, then they are assigned sequentially as the 4th–5th, 6th–7th elements, and so on within the same row. Finally, in rows with fewer pairs, the missing elements of the matrix are filled by repeating the last available pair of x_1 and x_2. In order to determine the possible clusters of the outputs we execute SVD again to obtain Eq. (12):

$$\mathbf{T}^O \overset{svd}{=} \mathbf{U}^O\mathbf{D}^O\left(\mathbf{V}^O\right)^T = \mathbf{U}^O\mathbf{S}^O. \tag{12}$$

  Here, we can disregard matrix $\mathbf{V}^O$ because we are looking to find dominant clusters along the first dimension only, which corresponds to an ordering of the output values.

  To find dominant or inverse dominant clusters in the outputs, we execute the CNO- or IRNO-type transformation to derive matrix $\mathbf{U}^{O,C}$ or $\mathbf{U}^{O,I}$. Finally, we have $\mathbf{T}^O$ as shown in Eq. (13):

$$\mathbf{T}^O = \mathbf{U}^{O,C}\mathbf{S}^O \quad \text{or} \quad \mathbf{T}^O = \mathbf{U}^{O,I}\mathbf{S}^O, \tag{13}$$

  based on which – using the pseudoinverse operator – we can obtain $\mathbf{S}^O$ as shown in Eq. (14):

$$\mathbf{S}^O = \left(\mathbf{U}^{O,C}\right)^+\mathbf{T}^O \quad \text{or} \quad \mathbf{S}^O = \left(\mathbf{U}^{O,I}\right)^+\mathbf{T}^O \tag{14}$$

  To align the number of clusters with the number of consequent values in $\mathbf{S}$ (that will be derived in the next step), the principal dominance can be identified by discarding negligible non-zero singular values in Eq. (12) and subsequently repeating the procedures outlined in Eq. (13) and Eq. (14).

- **Step 4: Determination of the Fuzzy Rules**

  Rationale: With both the antecedent matrices $(\mathbf{U}^C, \mathbf{V}^C)$, and the output structure $(\mathbf{T}^O)$ in place, the consequent matrix $\mathbf{S}^C$ is derived as the least-squares pseudo-inverse solution that best reconstructs the empirical data surface using the identified antecedent partition. The pseudo-inverse is used rather than a plain matrix inverse because the antecedent matrices are generally non-square. The result is a compact SCO matrix whose (i, j) entry directly quantifies the typical output level when input dimension 1 is in fuzzy set i and input dimension 2 is in fuzzy set j, enabling direct linguistic interpretation.

  From Step 2, we had the expression shown in Eq. (15):

$$\mathbf{T} \approx \mathbf{U}^C\mathbf{S}^C\left(\mathbf{V}^C\right)^T \tag{15}$$

  Once we have $\mathbf{U}^C$ and $\mathbf{V}^C$ as the discretised variant of the antecedent fuzzy sets (or indeed their IRNO counterparts) we can derive the fuzzy rule structure as shown in Eq. (16):

$$\mathbf{S}^C = \left(\mathbf{U}^C\right)^+\mathbf{T}\left(\left(\mathbf{V}^C\right)^T\right)^+. \tag{16}$$

  where either or both of $\mathbf{U}^C$ and $\mathbf{V}^C$ could be replaced by their IRNO counterparts.

  Whenever CNO matrices are used, the fuzzy rules will take the form of Eq. (17):

$$\text{IF} \quad x_1(t) \text{ is } A_{1,i} \quad \text{AND} \quad x_2(t) \text{ is } A_{2,j} \quad \text{THEN} \quad [\mathbf{S}^C]_{i,j}. \tag{17}$$

  This is because for every pair of (i, j) coordinates, only one column in the corresponding i-th and j-th rows of $\mathbf{U}^C$ and $\mathbf{V}^C$ will be substantially different from 0 (and close to 1). Regardless, the consequents can be calculated in the IRNO case, too, as a convex combination of the values within the rows and columns of matrix $\mathbf{S}^C$. Because

of its importance in interpreting relevant output values, this matrix will be referred to as the **SCO matrix** (the S-convex matrix) in further analyses.

Based on the clusters uncovered in step 3, the consequent values can also be associated with specific linguistic terms, such as “low”, “medium” or “high”. Importantly, these will be derived based on the data inside the $\mathbf{T}^O$ matrix.

4.1. Validation of the derived fuzzy rule sets

The rule sets obtained using the proposed framework can be validated from multiple perspectives.

First, in terms of external data consistency, the representativeness and quality of the measurement data can be assessed using conventional statistical diagnostics (e.g., sampling considerations, noise analysis). This ensures that the dataset itself is sound.

Second, in terms of fidelity or internal consistency, the ETSFM model is expected to approximate the empirical decision surface and should reproduce observed behavior. Internal consistency, then, is quantified by the agreement between model-based reconstructions and the empirical response grid. This means using reconstruction accuracy measures, such as R^2 scores for varying ranks; higher ranks increase precision, while more compact models retain interpretability.

Third, when it comes to stability, or robustness of the discovered structure, a key objective is to discover stable antecedent and consequent fuzzy sets. Here, we can evaluate whether the extracted rule backbone persists across model ranks and antecedent-shaping variants.

Classical significance tests (e.g., p-values, confidence intervals) could be applied in a preliminary stage to validate that the dataset reflects systematic behavior rather than noise; however, they do not evaluate whether the resulting fuzzy rule base faithfully captures the functional structure of the decision process. Therefore, in later parts of the paper we will focus on the latter aspects.

4.2. Computational complexity of the ETSFM transformation

One of the more computationally intensive aspects of the proposed method is the singular value decomposition (SVD). Given the importance of the CNO transformation in the method, an assessment of its complexity is also necessary.

For our derivations below, we assume that the number of discretization points in our 2-dimensional state space are M_1 and M_2 in the two dimensions, respectively. We also assume that the rank (i.e. number of weighting functions kept) in the two dimensions are R_1 and R_2, respectively.

4.2.1. Complexity of SVD

Note that the SVD algorithm, carried out over a matrix of size m-by-n, has a complexity of $O(mn \cdot \min(m,n))$ (Vasudevan & Ramakrishna, 2017). Thus, in our case we have the complexity shown in Eq. (18):

$$\text{complexity } = O(M_1 M_2 \cdot min(M_1, M_2)) \tag{18}$$

If the number of discretization points are comparable in the two dimensions, then the order of magnitude of growth can be maximized by a third-degree polynomial, which will be acceptable in a wide range of practical scenarios.

4.2.2. Complexity of the CNO transformation

The CNO transformation proposed in the literature (see e.g. Baranyi et al., 2017) performs a geometric embedding of the data into a simplex structure via polar coordinate optimization, so that the resulting weighting functions behave like fuzzy membership functions that preserve convex structure.

From a complexity perspective, the CNO transformation scales linearly with the number of discretization points but includes a polynomial-in-rank iterative search, therefore its runtime grows quickly with the rank due to repeated non-linear optimizations. Given that a part of these operations are internal to the implementation of the optimizer (such as Nelder-Mead), we estimate the complexity as in Eq. (19):

$$\text{complexity } = O(\sum_{i=1}^{2} M_i^2 R_i + N_{calls,i} N_{fev,i} M_i R_i + M_i R_i^3 + R_i^4) \tag{19}$$

where $N_{calls,i}$ is the number of calls to the Nelder-Mead optimizer, and $N_{fev,i}$ is the number of function evaluations it performs.

Given that we have d + 1 simplex points, where $d = R_i(R_i - 2)$, and given that Nelder-Mead performs a quadratic number of function evaluations in the number of parameters, we can estimate $N_{fev_i} = O(R_i^4)$.

At the same time, $N_{calls,i}$ is a hyperparameter on the order of 200 in our current implementation. This parameter specifies the number of times minimization is performed based on differently sampled simplex vertices and different perturbations so as to escape local minima.

Overall, the expected complexity would be on the order shown in Eq. (20):

$$\text{complexity } = O(\sum_{i=1}^{2} M_i^2 R_i + 200 M_i R_i^5 + M_i R_i^3 + R_i^4) \tag{20}$$

making $M_i^2 R_i + 200 * M_i R_i^5$ the dominating term. Depending on the rank of the weighting matrices and the number of data points, the relative weight of these two terms would be different.

4.2.3. Putting it all together

Based on the above, we have that the complexity of the proposed Extended TS Fuzzy Model Transformation is shown in Eq. (21):

$$\text{complexity } = O(M_1 M_2 \cdot min(M_1, M_2) + \sum_{i=1}^{2} M_i^2 R_i + 200 * M_i R_i^5) \tag{21}$$

If only a small number of singular values are kept compared to the number of discretization points, R_1 (R_2) will be negligible compared to M_1 (M_2) points and can be regarded as small constants. In this case, the complexity of the SVD algorithm would be the most costly part of the implementation. This is the general case, as the ranks to be kept are often chosen as small values.

However, in our application (Section 6), $M_1 = M_2 = 19$ discretization points, in which case, even when $R_1 = R_2 = 2$, the Nelder-Mead optimization steps dominate the computational cost. Substituting $M_1 = M_2 = 19$ and $R_1 = R_2 = 2$ into Eq. (21), the SVD term evaluates to $O(19 * 19 * 19) \approx O(6859)$, while the dominant CNO term evaluates to $O(2 * 19^2 * 2 + 200 * 19 * 2^5) \approx O(1.2 * 10^5)$, confirming that CNO optimization is the computational bottleneck in this regime. In practice, the entire ETSFM pipeline runs in under two minutes on a standard laptop for the 2-by-2 model, making it viable for offline analysis.

5. Application of the methodology in gray-box dynamic control assessment

In this section, we describe a framework for the application of the proposed Extended TS Fuzzy Model Transformation in the context of interpreting human behavior in gray-box dynamic control assessment.

5.1. Data acquisition framework

Within the framework, the goal is to understand how users interact with the gray-box system using the following concepts:

Definition 2 (State space and acceptable region)**.** *The system state is described by a set of parameters* $\mathbf{x}$*, and the possible values of that vector is referred to as the* ***state space****. The system state is said to be within the* ***acceptable region*** *if there exist a lower bound* $\mathbf{b}_l$ *and an upper bound* $\mathbf{b}_u$ *such that* $\mathbf{b}_l \leq \mathbf{x} \leq \mathbf{b}_u$.

Among the core concepts of the framework, the concepts of *trials*, *steps* and *rounds* are key in the data acquisition process:

Definition 3 (Trial, pre-update and post-update state). *A **trial** is a single decision instance in which the system starts from a **pre-update state** $\mathbf{x}_{pre}$, and the user has a single opportunity to perturb the system in some direction. Following the perturbation, the user finds the system in a **post-update state** $\mathbf{x}_{post}$.*

Definition 4 (Step and maximum attempts). *Within a single **step**, the user is given at most **maximum attempts** number of trials to perturb the system (by providing control inputs) in such a way that the post-update state $\mathbf{x}_{post}$ is within the acceptable region. If the user manages this, the step is considered* successful*; otherwise it is* unsuccessful.

Definition 5 (Round and required steps). *A **round** consists of some required number of consecutive successful steps (**required steps**). If any step within the sequence is unsuccessful, the round fails; otherwise it is successful.*

Definition 6 (Session end and required rounds). *A session terminates after some required number of successful rounds (**required rounds**), or upon reaching a time limit, whichever occurs first.*

5.2. Qualitative concepts describing user performance

To analyze user performance, we introduce the following qualitative notions:

Definition 7 (Self-stabilization region). *A local state-space zone where small variations in the state can be compensated by small (or zero) control adjustments to maintain or restore the state within the acceptable region; outside this zone, comparable variations require disproportionately larger inputs.*

Definition 8 (Blind spot). *A state-space region in which repeated trials consistently yield no state in the acceptable region within the maximum-attempt limit.*

6. Experimental framework and objectives

This section outlines the conceptual basis of the experimental setup designed to investigate human decision-making in the control of a balancing dynamic system. The study's primary aim is to identify the cognitive mechanisms involved in manual stabilization strategies by analyzing behavioral responses under varying stages of system instability.

6.1. Experimental task description

Participants are tasked with stabilizing a model that – at its base level – functions as an inverted pendulum mounted on a movable cart. Instead of direct physical interaction, participants influence the system by inputting numerical force values via a keyboard. These inputs simulate external forces acting on the cart to maintain the pendulum's balance.

Although this is a well-known benchmark problem from control theory, a key feature of the experiment (besides manual numerical input) is that participants are not given any information about the underlying dynamics, and – as detailed below – the visual representation of the system is transformed and simplified significantly so as to not be recognizable as a pendulum on a cart. Thus, as shown on Fig. 2:

- Participants view the angular position of the bob at the tip of the pendulum as a projection on a single (horizontal or vertical) axis.
- The visualization does not include the cart, nor any part of the pendulum apart from the bob at its tip.
- The angular velocity of the bob ($\dot{\theta}$) is indicated by a red arrow, with length proportional to magnitude.
- As shown on Fig. 2, the lines labeled B mark the boundaries of the measurement area, while the lines labeled S define the acceptable region. However, the primary objective is to maintain the bob as close as possible to the central line labeled C. The measurement area, defined by lines B, corresponds to the boundaries of $\theta \in [-60°, 60°]$. The acceptable region, bounded by lines S, corresponds to $\theta \in [-25°, 25°]$.
- No mechanical model, animation, or time-based simulation is shown; only the immediate result of the participant's input is visualized.

The absence of animation and continuous feedback transforms the dynamic system into a series of discrete, symbolic updates. This intentional design turns the computer into a cognitive filter, removing sensorimotor involvement and enforcing a purely abstract, screen-based reasoning strategy. The computer thus plays an active role in reshaping the participant's engagement with the task. To help provide an intuitive understanding of the task, participants are given the following metaphor: "*Imagine a mountain where the central white line is the ridge, the surrounding white area represents the summit, and the blue margins indicate the base. Your goal is to keep the ball from rolling off the mountain.*"

Further implementation details regarding the measurement process, input handling, timing, and success criteria are described in Section 6.4.1.

6.2. Second-order dynamic representation of the system

In the following, we define the equations of motion of the second-order balancing-type dynamic system utilized in the experimental setup.

The left image of Fig. 2 provides a mechanical illustration of the inverted pendulum. The quasi Linear Parameter Varying (qLPV) state-space model structure is shown in Eq. (22):

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \mathbf{S}(\mathbf{p}(t))\begin{bmatrix}x_1(t)\\ x_2(t)\\ u(t)\end{bmatrix}, \quad \text{where} \quad \mathbf{p}(t) = \begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix}, \tag{22}$$

where $\mathbf{S}(\mathbf{p}(t))$ is defined as in Eq. (23):

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ -\frac{g}{l}\frac{\sin(x_1)}{x_1} & 0 & -\frac{1}{ml^2}\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\\ u(t)\end{bmatrix}, \tag{23}$$

where $x_1(t)$ denotes the angle $\theta \in [-60\pi/180, 60\pi/180]$ (in radians) of the pendulum from the vertical direction, and $x_2(t)$ is the angular velocity $\dot{\theta} \in [-50\pi/180, 50\pi/180]$. The input $\mathbf{u}(t)$ represents the applied torque, measured in Newton-meters (Nm). Further, $g = 9.8\frac{m}{s^2}$ is the gravity constant, and in our experiments, $l = 1m$ is the length of the pendulum and $m = 5kg$ is the mass of the pendulum.

6.3. Reference controller for the system

To provide a consistent benchmark for interpreting the human control strategies described in Sections 6.1–6.7, we derived a mathematically optimized reference controller using an LMI-based Parallel Distributed Compensation (PDC) design framework (Tanaka & Wang, 2001). This controller serves as a basis for quantitative and qualitative comparisons with the human intervention patterns, which are discussed in detail in Section 7.

This framework assumes that the TS fuzzy model (Eq. (2)) of $\mathbf{S}(\mathbf{p})$ is provided, and the goal is to control $\mathbf{x} \to \mathbf{0}$. It then derives the controller, defined in Eq. (22), in TS fuzzy model form as in Eq. (24):

$$\mathbf{u} = -(\mathbf{F}(\mathbf{p}))\mathbf{x} \tag{24}$$

where $\mathbf{F}(\mathbf{p})$ is defined as in Eq. (25):

$$\mathbf{F}(\mathbf{p}) = \sum_{i_1=1}^{I_1}\sum_{i_2}^{I_2} w_{1,i_1}(p_1)w_{2,i_2}(p_2)\mathbf{F}_{i_1,i_2}. \tag{25}$$

where the antecedent system $w_{1,i}(p_1)$ and $w_{2,i_2}(p_2)$ of the controller is inherited from the TS fuzzy model (Eq. (2)) and the vertices $\mathbf{F}_{i_1,i_2}$ are derived by LMIs from the vertices $\mathbf{S}_{i_1,i_2}$. There exist a wide variety of LMIs, each with distinct properties that aid in optimizing control performance (Scherer & Weiland, 2000).

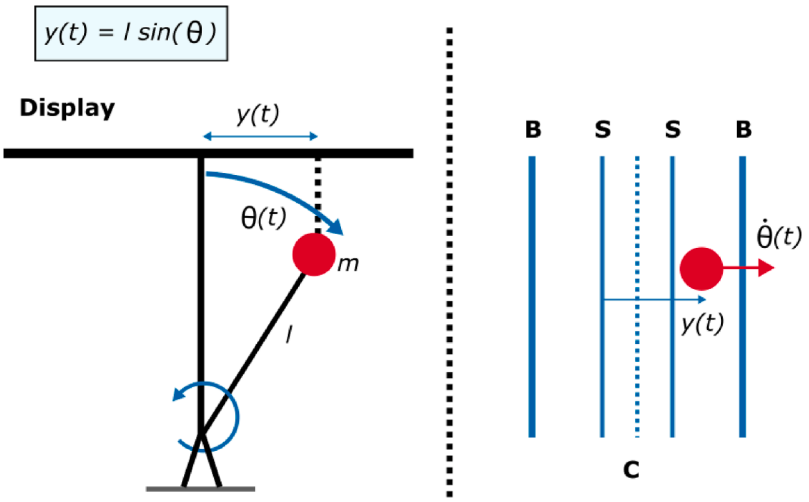

Fig. 2. The dynamic system (left) and its visualization (right).

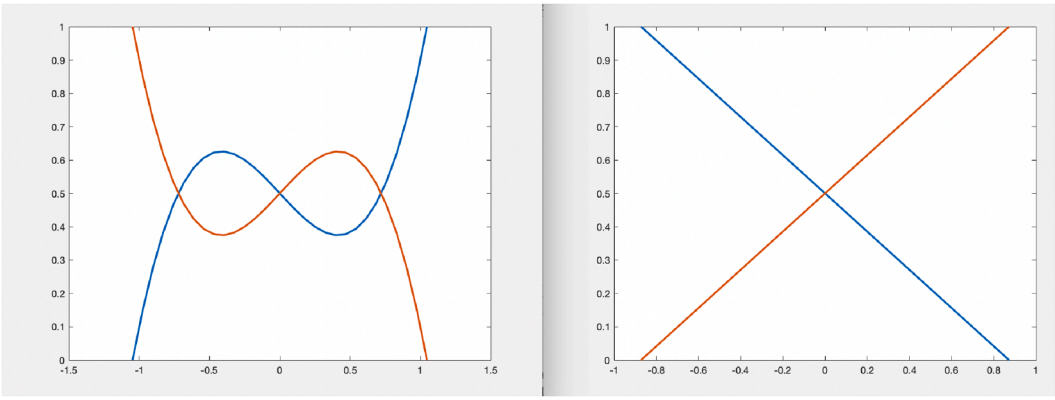

Fig. 3. Weighting functions of the reference controller per state space dimension.

The key conclusion of this section is that the controller relies on only 4 rules, and its antecedents exhibit symmetry — a natural property, given that the balancing pendulum system itself is inherently symmetric (Fig. 3).

6.4. Measurement procedure

In this subsection, we describe the implementation details, the participants and the instructions they received.

6.4.1. Implementation details

The dynamic system and participant interface are implemented on a personal laptop with a 15.6-inch screen for visualization. Participants enter input values via the keyboard into a designated text box on the screen. The direction of the input force is then selected by clicking on left- and right-oriented arrows (see Fig. 4).

Once the selected arrow is clicked, the next position and speed of the bob are immediately visualized. The updated position represents the system's state after a simulated process of 3 time steps, but this time progression exists only within the simulation and is not real-time. The participant receives the result instantly after clicking the arrows, without any experience of the time that is simulated. This also means that the bob's movement is not animated; instead, after the initial position and velocity are displayed and the input given, only the resulting position and velocity are immediately visualized. The combination of static visualization, text-based input, the absence of mechanical system animation, the lack of bob movement animation, and the omission of process

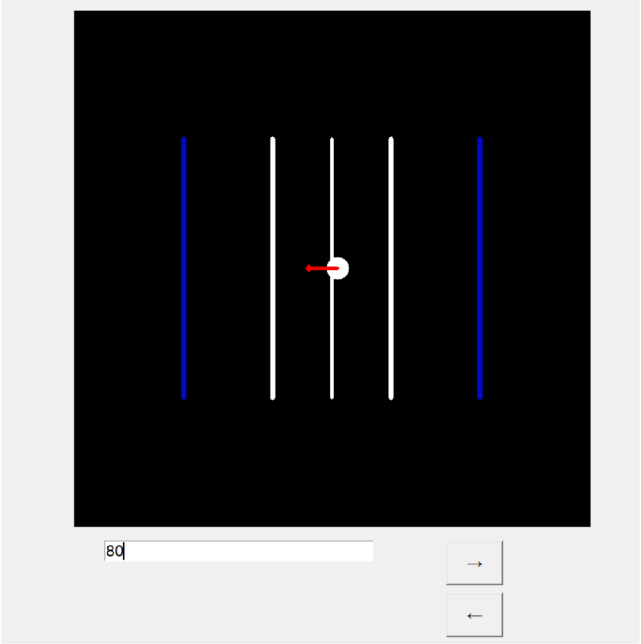

Fig. 4. Measurement environment – vertical case.

time information minimize the test participants' ability to perceive the system's dynamics directly, providing only abstract information.

In each session, defined as a sequence of steps with specified initial parameters, the supervisor initializes the system with a given angle and angular velocity. Participants enter numeric force values and select a direction using the minimal control panel.

After successful stabilization or ten unsuccessful attempts, the supervisor initiates the next step. This structure sets the stage for the procedural details described in the following subsection.

6.4.2. Test subjects and instructions

A total of 50 volunteers participate in the study. Of these, 5 participants (3 male, 2 female) belong to Generation Y, another 5 (3 male, 2 female) to Generation X, and the remaining 40 (30 male, 10 female) to Generation Z. The participants receive the task description below:

> *Dear Participant,*
>
> *In this task, you will see a point. The direction of the point's speed is indicated by an arrow.*
>
> *Your task is to keep the point within the white lines for 10 consecutive steps, across a total of 3 rounds. The edges of the playing field are marked by the two blue lines. You can set the speed value by entering a number in the empty rectangle, and then choose the intervention direction by clicking on the appropriate arrow. This will also start the movement of the point.*
>
> *You can control the interface using a mouse, touchpad, touchscreen, or keyboard. For mouse, touchpad, or touchscreen control, click on the control arrow to initiate the action. If controlling with the keyboard only, after entering the value, use the Tab key to switch to the arrows (1x Tab for upward direction, 2x Tab for downward direction). In this case, confirm the direction with Enter.*
>
> *The goal is to understand how different speeds result in different levels of change.*
>
> *If the point goes beyond the blue line, you have exited the playing field, and you can try again from the previous state.*
>
> *A round ends when you successfully keep the point within the white lines for 10 consecutive steps. After that, the next round begins. The game ends when you either successfully complete all 3 rounds or reach the 20-minute time limit."*

To help conceptualize the task, participants are also asked to imagine the mountain metaphor from Section 6.1.

7. Measurement results and fuzzy model based analyses

This section presents the findings of the experiment in three key dimensions. First, participant decision-making strategies are analyzed based on the magnitude and direction of applied intervention forces. Second, human control behavior is described using fuzzy IF-THEN rules derived from mathematical modeling. Third, the cognitive complexity of decisions is assessed through the analysis of rule weights and reaction domains, highlighting asymmetries and decision uncertainty.

Instead of statistical comparisons, the focus is on interpretability: identifying generalizable control heuristics and typical behavioral responses across varying system states. The fuzzy modeling approach enables the visualization of abstract decision zones and the structural patterns behind them. Key patterns are presented through increasing levels of model complexity—starting from minimal rule sets and extended to more refined representations.

Fig. 5 links the grid-structured $\mathbf{T}^O$ matrix to three interpretable outputs: rules, decision-region mapping, and a human-controller comparison.

7.1. Derivation of suitable antecedents

In this subsection, we describe how the proposed method is applied to the measurement data. During the procedure, the measured control inputs are assigned into predefined intervals and binned into the matrix $\mathbf{T}$. We then apply the proposed Extended TS Fuzzy Model Transformation to this matrix.

Based on singular value decomposition (SVD) of the matrix, the relative weight of each component (i.e., columns, or weighting functions in the weighting matrices) is considered. Columns with negligible contribution to system behaviour are excluded so as to reduce the number of weighting functions, hence fuzzy rules. This is often seen as an acceptable trade-off between complexity and interpretability.

7.1.1. Validation of rank-reduced models for internal consistency

Before proceeding, we want to ascertain that the rank-reduced systems still provide a good approximation to the original system. To this end, we reconstruct the data in the original grid-points using the rank-reduced system, and calculate the R^2 statistic between the original data and the reconstructed data.

Results obtained for systems of rank 2-by-2, rank 3-by-3 and rank 4-by-4 are as follows:

- R^2 (rank 2-by-2) $= 0.9835$
- R^2 (rank 3-by-3) $= 0.9929$
- R^2 (rank 4-by-4) $= 0.9952$

This shows that even in the case of the rank 2 system, which is the coarsest approximation among the three, over 98% of the original variance is kept in the reconstructed data. Although this metric would not be sensitive to output values being "swapped" between grid points, aside from this unlikely event it shows that the rank-reduced systems provide a useful and good approximation of the original data.

7.1.2. Dominant antecedent fuzzy sets in the 2-by-2 model

The most interpretable configuration is the 2-by-2 fuzzy model, which employs two antecedent fuzzy sets for the angular position and two for angular velocity. This model effectively captures the dominant behavioral tendencies observed during the experiment.

Fig. 6 visualizes the main components of human decision-making based on this model following the CNO transformation. Two distinct antecedent fuzzy sets (together with a third, normalizing set) emerge as dominant:

- In the first dimension (angle), the blue antecedent fuzzy set fires alone near $0°$, and together with the yellow antecedent fuzzy set for most angles. The red antecedent fuzzy set, by contrast, is most relevant near $40°$.
- In the second dimension (angular velocity), two major antecedent sets are identified – at around $20\ deg/\text{sec}$, and between -30 and $10\ deg/\text{sec}$ – while the third one becomes prominent around $-50\ deg/\text{sec}$.

These patterns demonstrate an asymmetric decision structure, where users respond differently to leftward and rightward deviations. Despite system symmetry, the applied control logic reflects cognitive preferences and directional biases.

7.1.3. Increasing the number of antecedents

Note that while extending the antecedent fuzzy set base to three or four (see Fig. 7 for the latter case) rules per input dimension increases the granularity of captured behaviors, the primary regions that are identified within the antecedent space as being relevant to the fuzzy rules are left unaltered, with the two dominant antecedent fuzzy sets consistently reappearing.

However, the additional antecedent fuzzy sets in Fig. 7 may reflect behavioral nuances that could stem from latent individual factors such

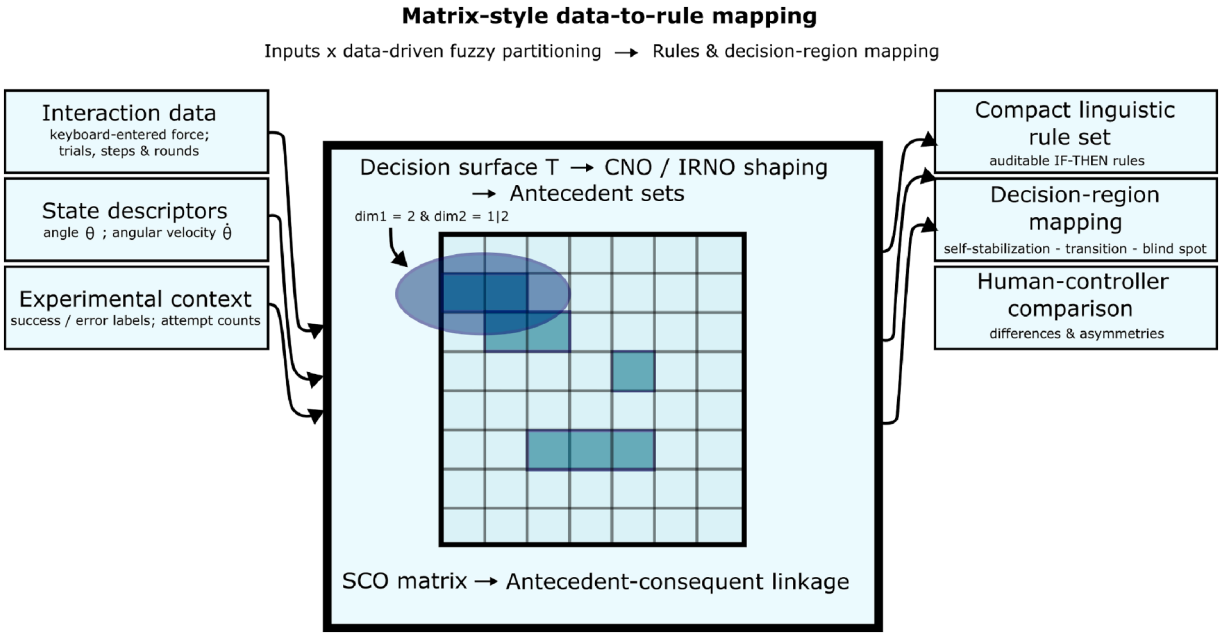

Fig. 5. Interpretable outputs: compact rules, region mapping, controller comparison.

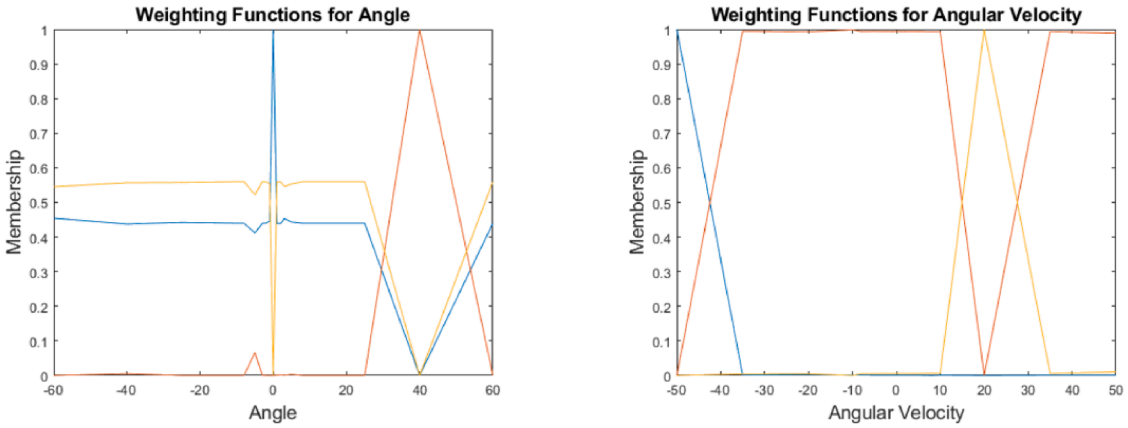

Fig. 6. CNO weighting functions for angle (left panel, horizontal axis in degrees) and angular velocity (right panel, horizontal axis in deg/s) in the 2-by-2 rank system. The vertical axis in both panels represents the membership weight in the range $[0, 1]$.

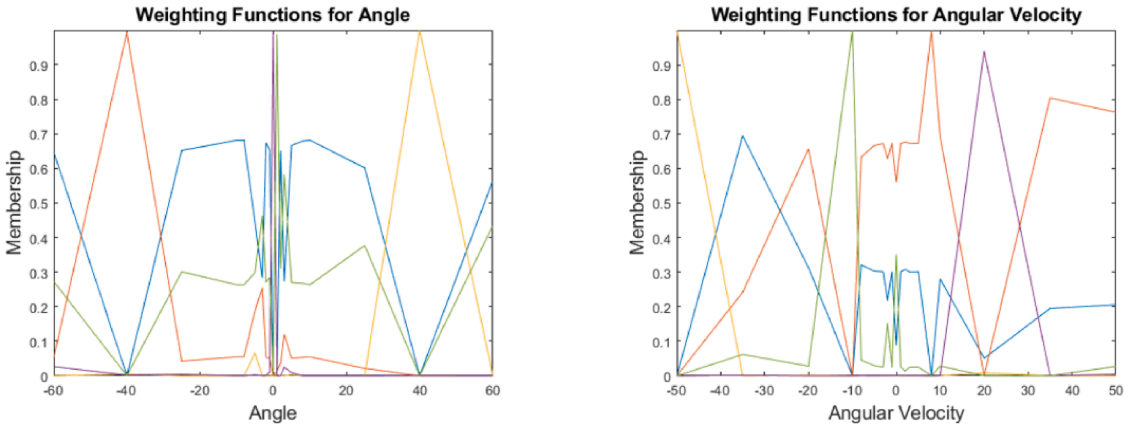

Fig. 7. CNO weighting functions for angle (left panel, horizontal axis in degrees) and angular velocity (right panel, horizontal axis in deg/s) in the 4-by-4 rank system. The vertical axis in both panels represents the membership weight in the range $[0, 1]$. The additional fuzzy sets (relative to the 2-by-2 model) refine boundary regions without altering the dominant antecedents.

as hand dominance (left- vs. right-handedness), impulsivity, or external distractions. These underlying influences are often difficult to isolate or explain explicitly.

For applications such as decision-support systems, it is essential to strike a balance between behavioural resolution and clarity. This balance is also emphasized in a previous work on user feedback modeling, where a low-complexity fuzzy antecedent fuzzy set base proves sufficient to capture key behavioral patterns in human-software evaluation tasks (Sudár et al., 2025). The current model applies this principle to dynamic control environments, demonstrating that core cognitive structures can be retained even in minimal models.

7.2. Derivation of fuzzy rules and identification of decision strategies

Based on the resulting SCO matrix, along with the supports and peaks of the fuzzy sets uncovered from the 2-by-2 system of weighting functions, we identify the following key rules:

$$\text{Force} = \begin{cases} \text{L \& negative}, & \text{if } -10 < \theta \leq 0 \& \dot{\theta} < -35 \\ \text{L \& negative}, & \text{if } \theta \approx 0 \& \dot{\theta} \in [-30, 10] \\ \text{L \& positive}, & \text{if } -10 < \theta < 0 \& \dot{\theta} \approx 20 \\ \text{L \& positive}, & \text{if } \theta > 30 \& \dot{\theta} \approx 20 \\ \text{M \& negative}, & \text{if } \theta > 30 \& \dot{\theta} < -35 \\ \text{M \& negative}, & \text{if } \theta > 30 \& \dot{\theta} \in [-30, 10] \\ \text{S \& negative}, & \text{if } -10 < \theta < 0 \& \dot{\theta} \in [-30, 10] \\ \text{S \& positive}, & \text{if } \theta \approx 0 \& \dot{\theta} \approx 20 \end{cases}$$

where L, M and S stand for “large”, “moderate” and “small”, respectively; $vel.$ and $dir.$ stand for “velocity” and “direction”; θ represents the angle and $\dot{\theta}$ represents the angular velocity of the bob.

These rules can be further simplified into the following key takeaways:

- IF bob is close to center & velocity is small THEN apply small force.
- IF bob is to the left & velocity is substantial THEN apply large force.
- IF bob is to the right & velocity is substantial THEN apply moderate force
- IF velocity is high THEN apply large force.

These conditional antecedent fuzzy sets mirror observed user behavior, demonstrating that decisions are not random but follow rules with a structured internal logic. Further, the fuzzy model effectively captures these structures, offering interpretable insights into human control strategies.

7.3. Cognitive load and decision difficulty across system states

In the next step, decision difficulty and cognitive load are categorized into different levels based on the number of attempts required to identify a suitable intervention force, along with participants’ subjective feedback. Although response times are also recorded, no significant differences are found across system states in the average time spent per intervention. However, scenarios with higher cognitive load typically result in more repeated attempts, leading to longer cumulative decision times.

Fig. 8 presents a matrix visualizing cognitive load intensity, where different colors indicate the relative difficulty of specific decision scenarios. Light green regions indicate low difficulty, where participants typically found the correct intervention force within one or two attempts. Yellow zones represent moderate difficulty with three to four attempts, orange indicates higher difficulty requiring five to six attempts, and red marks high difficulty regions requiring more than seven attempts. Brown areas represent decision blind spots – states where participants consistently failed to find any successful control input, despite repeated attempts.

These results are consistent with the findings presented in Fig. 9, which shows the error rates of human decision-making. The spatial distribution of errors closely matches the difficulty levels identified in the cognitive load matrix (Fig. 8).

8. Discussions

In this section, we discuss the advantages of the proposed ETSFM approach, and the conclusions drawn based on its application within the second-order gray-box dynamic control experiment. We also highlight some of the caveats and limitations that could help motivate future work.

8.1. Novelty and strengths of ETSFM

Understanding human control in dynamic tasks often comes at a trade-off: parametric models risk misspecification, while black-box learners trade accuracy for opacity. In this paper, we propose a TS fuzzy model transformation based approach that allows for the structural identification of raw measurement data and the transformation of the identified model into interpretable fuzzy rules.

A key novelty of the ETSFM approach is that it is non-parametric and data-driven, mapping grid-structured interaction data directly to an auditable fuzzy rule base without pre-defined membership templates or hand-crafted rules, preserving the measurement-implied structure.

With as few as 2×2 antecedents, the model reveals stable and interpretable regimes (self-stabilization, transition-driven uncertainty, directional biases) while remaining sufficiently compact for inspection and communication.

From a practical standpoint, the running time is primarily driven by the SVD-based decomposition step. Under low-rank (e.g., 2×2) and high-discretization settings (e.g., upwards of $1,000$ discretization points), the CNO/IRNO decomposition step is negligible in comparison, yielding overall polynomial scaling (fifth-order at most in the rank, but second-order in the measurement resolution), which makes it a practical and viable option for offline analyses and periodic re-fitting.

The method benefits from grid-like coverage of the state space; sparse or biased sampling can under-represent transition regions. This risk can be mitigated by fidelity/stability checks and, where necessary, targeted data collection in boundary zones.

ETSFM can be integrated as an offline analytics module alongside existing decision workflows, providing explainable summaries of decision regions without modifying the operational control loops.

8.2. Results on second-order gray-box dynamic control

Through our experiment, we address the accuracy-interpretability tradeoff by recovering an auditable, low-complexity rule structure of human decision logic directly from empirical responses and contrasting it with an optimized, symmetric controller. Our guiding question is whether an interpretable fuzzy rule base can capture the dominant regimes of behavior and where human control systematically diverges from optimal control.

A compact 2-by-2 fuzzy model (two antecedent sets per input) captures the dominant structure of the empirical decision surface while remaining easy to read and audit. The SCO matrix indicates a small number of consequential force levels (Small/Moderate/Large) consistent with observed intervention patterns, suggesting participants follow a rule-like logic rather than behaving randomly. Cognitive load clusters in transition regions where position and velocity cues conflict, yielding wider force dispersion and higher errors. We also observe decision blind spots adjacent to easy-to-control regions, consistent with threshold-like internal representations that produce sharp performance drops.

Turning to the directional structure of these results, along the angular-velocity axis we consistently find two dominant antecedent sets – one near $0\ deg/s$ and another around 20–50 deg/s. Near $0\ deg/s$ and

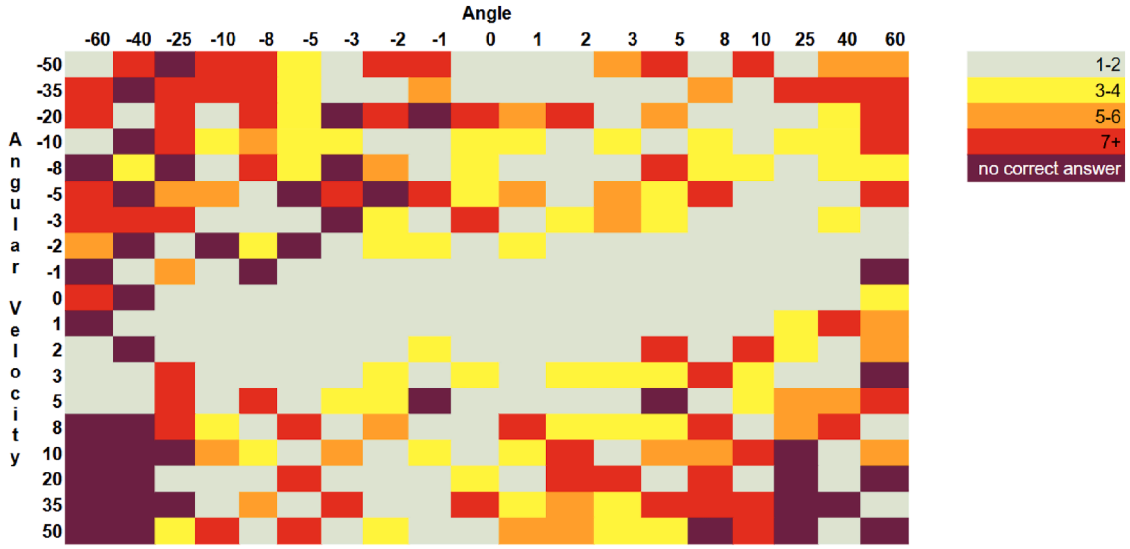

Fig. 8. Intensity of cognitive load based on number of trials needed to complete a step for any given pair of angle (horizontal axis, degrees) and angular velocity (vertical axis, deg/s) values. Color legend: light green = 1–2 attempts (low difficulty); yellow = 3–4 attempts; orange = 5–6 attempts; red > 7 attempts (high difficulty); brown = decision blind spot (no successful input found within attempt limit). The figure confirms that extreme angles and / or angular velocities are more difficult to handle. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

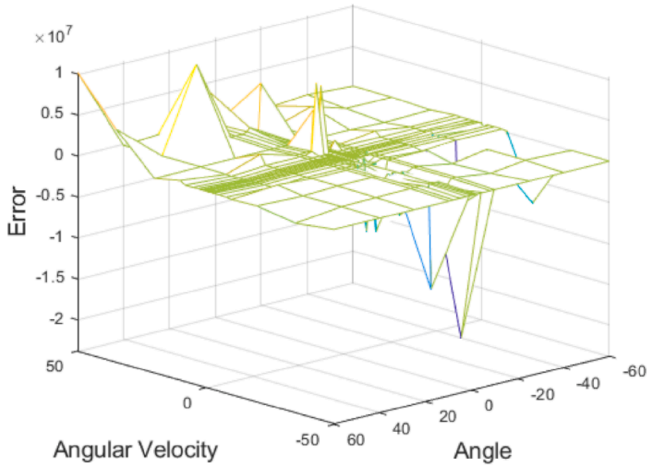

Fig. 9. Error rate of human decision-making across the state space (horizontal axis: angle in degrees; vertical axis: angular velocity in deg/s). Each cell represents the aggregated magnitude (RMSE) of the force values entered by participants in the given state. The spatial distribution of errors mirrors the cognitive load intensity shown in Fig. 8, confirming that high-difficulty regions coincide with high error rates.

≈ 0° displacement, participants have a *safe exploration window* to probe system response, which encourages high-variability, small adjustments. By contrast, at 20–50 deg/s the state departs from stability much faster, and participants may feel they have less time to decide, leading to more decisive interventions. In combination with the position antecedents (including recent angle history), this yields an asymmetric decision structure: states mirrored across the origin $(\theta, \dot{\theta}) \rightarrow (-\theta, -\dot{\theta})$ are not treated equivalently. Practically, the same kinematic "shape" on the left vs. right side triggers different control choices, consistent with directional biases and cognitive preferences. Increasing model rank preserves these core regimes, and the added complexity mostly refines edge cases near regime boundaries without altering the core underlying patterns.

Comparing these human patterns with the optimized reference controller, the reference controller employs four rules with symmetric antecedents and delivers consistent regulation across the state space, including peripheral regions where humans struggle. Human behavior overlaps with the controller's self-stabilization region but diverges in transition zones, showing fragmentation and asymmetry-signatures of heuristic, context-dependent strategies under uncertainty.

To quantify this divergence, the error surface in Fig. 9 reveals that human control errors (RMSE) concentrate in the peripheral and transition regions of the state space, whereas the optimized controller maintains near-zero error throughout. The SCO matrices further expose the asymmetry: where the controller assigns symmetric force levels to mirrored states $(\theta, \dot{\theta})$ and $(-\theta, -\dot{\theta})$, human participants systematically apply different force magnitudes – for instance, states involving leftward deviations at high angular velocities elicit larger interventions than their rightward counterparts. This directional bias, absent from the symmetric controller, points to a cognitive rather than mechanical origin. In terms of robustness, the controller's performance is invariant across the state space by construction, while human performance degrades sharply at the boundaries of the self-stabilization region, as evidenced by the abrupt transition from low-difficulty (1–2 attempts) to blind-spot states (no successful input found) visible in Fig. 8.

From the perspective of cognitive science, the patterns uncovered by ETSFM align with several well-established theoretical frameworks. The self-stabilization regions, where small or zero inputs suffice, are consistent with the notion of "satisficing" in bounded rationality (Simon, 1956): participants appear to adopt a "good enough" strategy when the system is near equilibrium, rather than optimizing their inputs. The transition zones, where cognitive load and error rates spike, correspond to regimes where the dual-process model of cognition (Kahneman, 2011) predicts a shift from fast, intuitive System I responses to slower, deliberate System II reasoning – a shift that is cognitively costly and error-prone. The directional asymmetry – participants treating kinematically mirrored states differently – is consistent with the well-documented phenomenon of pseudoneglect (Jewell & McCourt, 2000): even neurologically healthy individuals show a systematic leftward bias in spatial attention tasks, attributed to right-hemisphere dominance in visuospatial processing. In the present task, such an attentional asymmetry could translate into different perceptual thresholds for leftward vs. rightward deviations and, consequently, different intervention magnitudes. A complementary explanation is motor lateralization: however, the absence of handedness data prevents a definitive attribution, and future stratified studies will be needed to disentangle attentional from motor sources of

the bias. While these connections are suggestive rather than conclusive, they demonstrate the ETSFM output is rich enough to support interdisciplinary interpretation and hypothesis generation.

8.3. Caveats and limitations

Several caveats and limitations should be noted. A current limitation of the ETSFM framework is that while it enables structural fidelity to be validated (e.g. through R^2 scores), as well as stability to be analyzed (e.g. through the use of differently constrained, CNO/IRNO weighting functions and different degrees of rank reduction); it does not currently focus on defuzzification of generated rules to generate inferences in a way that could be validated. In the current study, structural fidelity is validated through R^2 scores, and the stability of rules is validated through a grid search over different rank decompositions and CNO vs. IRNO normalization. This approach is viable in a practical sense due to the low dimensionality of the rank-by-normalization-type grid over which the search is carried out. It is also clear that the high reconstruction accuracy of the SVD-CNO/IRNO steps guarantees a high degree of precision of the rules – at least over the grid coordinates. Regardless, the methodology could benefit from a defuzzification extension that could help disambiguate rules that are potentially in conflict, and to perform inference between the antecedent sets of the rules in a way that could be validated.

With respect to the cognitive science aspects of the study, it should be noted that we intentionally filtered out sensorimotor cues and real-time dynamic feedback to prioritize high-level cognition. While this is a conceptual design decision, it limits direct generalizability to continuous, embodied control. Extending the study to real-time feedback tasks would increase the robustness of the conclusions and should be a part of future work.

It should also be noted that while the results of the study point to certain left-right asymmetries in control decisions, they do not extend to a deeper analysis on whether this could relate to e.g. hand-dominance or other individual traits. Future studies might adopt stratified sampling (e.g. handedness labels), mirrored mappings and other group-wise ETSFM analyses to quantify symmetry effects. This would also entail selecting a larger and more balanced population of participants. While the sample size of 50 used in the current study does allow for the ETSFM method to be validated, and for initial conclusions to be drawn, taking measurement data from a larger sample size is not the core focus of this study.

8.4. Future potential applications

In terms of potential applications, the framework could extend to a wide range of fields, including business, economics and AI. In business and economics, the need to make financial and managerial decisions under uncertainty is a common challenge that often prompts actors to rely on heuristics and pattern recognition instead of exclusively relying on quantitative models (Tversky & Kahneman, 1974). Analogous to our participants' intuitive force adjustments, traders and managers often balance historical patterns with present signals in real time (Lo, 2004). Given access to interaction logs with state-action pairs and adequate state-space coverage, ETSFM could be used to derive decision-region maps and a compact, auditable rule base that make directional biases and uncertainty-sensitive regimes explicit for supervisory analyses. Similarly, in AI applications such as autonomous or semi-autonomous driving, given access to state-action logs from simulation or test fleets, ETSFM could be used to derive decision-region maps and a compact, auditable fuzzy rule base that make deviations from a chosen benchmark (e.g., human patterns or a reference policy) explicit in transition-heavy scenarios. Such maps would highlight outlier regimes, directional asymmetries, and shifted intervention thresholds, guiding targeted data collection and human-machine interaction refinement. The approach could also support personalization by extracting user-specific rule sets for interface tuning.

9. Conclusions

This paper presents the Extended TS Fuzzy Model (ETSFM) transformation, a non-parametric, data-driven framework that converts structured interaction data into compact, linguistically labeled fuzzy IF–THEN rules. The aim is to characterize human decision logic in black-box or gray-box dynamic environments while emphasizing structure discovery and interpretability over predictive optimization.

The ETSFM framework is validated through an application scenario involving a second-order dynamic system that, unknown to users, exhibits characteristics of an inverted pendulum on a movable cart. Participants interact with a graphical user interface that intentionally abstracts away visual and dynamic cues. User inputs are collected as discrete reactions to static snapshots, eliminating any sense of a continuously unfolding simulation. In this setting, ETSFM successfully produces a low-rank fuzzy rule base that captures dominant decision behaviors at a high structural accuracy.

Clear differences emerge between human strategies and a mathematically precise reference controller. Human participants display directional asymmetries and blind spot behaviors, especially in transition regions, whereas the controller, defined by four symmetric rules, maintains consistent and balanced regulation. These results indicate heuristic, context-dependent human control patterns and illustrate how a compact, auditable rule base can make such differences explicit.

Future work should extend the framework to real-time tasks, conduct larger and stratified participant studies to investigate individual factors such as handedness, and explore training protocols that align human strategies with reference-controller behavior in transitional regimes. More generally, a defuzzification extension could enable ETSFM to infer new data points, supporting further validation and predictive modeling of human decision-making.

In summary, the theoretical contribution of the paper is a principled, non-parametric workflow that extends the TS fuzzy model transformation from analytically defined systems to raw behavioral data, producing auditable linguistic rules without pre-defined membership shapes. The practical contribution is a first systematic fuzzy rule based characterization of human control behavior in a gray-box dynamic task, yielding actionable insights into where human decision-making succeeds, where it fails, and how it structurally differs from optimal control. These contributions position ETSFM as a transferable tool for any domain where interpretable modeling of human decisions from interaction logs is desired – including process control, driver assistance, and human-AI collaboration.

Funding

This research received no external funding.

Ethics approval

Ethical approval for the research involved in this study was granted by Corvinus University of Budapest ref no: KRH/343/2024.

Consent to participate

Each participant read and accepted the declaration of consent before the measurements.

Consent for publication

Each participant read and accepted the declaration of consent before the measurements and accepted that the data provided can be used and published in scientific form.

Code availability

A Python implementation of the ETSFM pipeline can be found in the public repository at: https://github.com/csapoadam/interpretable-modeling-dynamic-stabilization/ Alternatively, the TP modeling toolbox for Matlab can be accessed via the Wikipedia page: https://en.wikipedia.org/wiki/TP_model_transformation_in_control_theory.

Generative AI was used only for language polishing

No AI system was used for study design, data collection, data analysis, or interpretation of results.

CRediT authorship contribution statement

Ildikó Horváth: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing; **Anna Sudár:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing; **Ádám B. Csapó:** Conceptualization, Software, Validation, Writing – original draft, Writing – review & editing.

Data availability

The anonymized measurement data can be found in the public repository at: https://github.com/csapoadam/interpretable-modeling-dynamic-stabilization/

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by the National Research, Development and Innovation Office (NKFIH) under grant number 2024-1.2.3-HU-RIZONT-2024-00030.

References

Angelov, P., & Yager, R. (2011). Simplified fuzzy rule-based systems using non-parametric antecedents and relative data density. In *2011 IEEE workshop on evolving and adaptive intelligent systems (EAIS)* (pp. 62–69). IEEE.

Baranyi, P. (2004). TP model transformation as a way to LMI-based controller design. *IEEE Transactions on Industrial Electronics*, *51*(2), 387–400.

Baranyi, P. (2014). The generalized TP model transformation for T–S fuzzy model manipulation and generalized stability verification. *IEEE Transactions on Fuzzy Systems*, *22*(4), 934–948. https://doi.org/10.1109/TFUZZ.2013.2278982

Baranyi, P. (2016). TP-model transformation based control design frameworks. In *Control engineering*. Springer. https://doi.org/10.1007/978-3-319-19605-3

Baranyi, P. (2020a). Extracting LPV and qLPV structures from state-space functions: A TP model transformation based framework. *IEEE Transactions on Fuzzy Systems*, *28*(3), 499–509. https://doi.org/10.1109/TFUZZ.2019.2908770

Baranyi, P. (2020b). How to vary the input space of a T–S fuzzy model: A TP model transformation-based approach. *IEEE Transactions on Fuzzy Systems*, *30*(2), 345–356.

Baranyi, P. (2023a). Dual-control-design TP and TS fuzzy model transformation based control optimisation and design. In *Topics in intelligent engineering and informatics*, Vol. 17. Springer. https://doi.org/10.1007/978-3-031-44575-0

Baranyi, P. (2023b). Transition between TS fuzzy models and the associated convex hulls by TS fuzzy model transformation. *IEEE Transactions on Fuzzy Systems*, *32*(4), 2272–2282.

Baranyi, P. (2024). Relaxed TS fuzzy model transformation to improve the approximation accuracy/complexity tradeoff and relax the computation complexity. *IEEE Transactions on Fuzzy Systems*, *32*(9), 5237–5247.

Baranyi, P., Yam, Y., & Várlaki, P. (2017). Tensor product model transformation in polytopic model-based control. In *Automation and control engineering*. CRC Press, Taylor & Francis Group. ISBN 9781138077782.

Beltrami, E. (1873). Sulle funzioni bilineari. *Giornale di Matematiche ad Uso degli Studenti Delle Universita*, *11*(4), 98–106.

Chiu, S. (1997). Extracting fuzzy rules from data for function approximation and pattern classification. *Fuzzy information engineering: A guided tour of applications*, *9*, 1–10

Epstein, J. M. (2012). Generative social science: Studies in agent-based computational modeling. Princeton University Press.

Gonzalez, C. (2022). Learning and dynamic decision making. *Topics in Cognitive Science*, *14*(1), 14–30.

Gu, X., & Angelov, P. P. (2020). Highly interpretable hierarchical deep rule-based classifier. *Applied Soft Computing*, *92*, 106310.

Guillaume, S. (2001). Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE Transactions on Fuzzy Systems*, *9*(3), 426–443.

Ishteva, M., De Lathauwer, L., Absil, P. A., & Van Huffel, S. (2008). Dimensionality reduction for higher-order tensors: Algorithms and applications. *International Journal of Pure and Applied Mathematics*, *42*(3), 337–353.

Jang, J. S. R. (1993). ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*(3), 665–685. https://doi.org/10.1109/21.256541

Jewell, G., & McCourt, M. E. (2000). Pseudoneglect: A review and meta-analysis of performance factors in line bisection tasks. *Neuropsychologia*, *38*(1), 93–110. https://doi.org/10.1016/S0028-3932(99)00045-7

Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference, and consciousness. Harvard University Press.

Kahneman, D. (2011). Thinking, fast and slow. New York: Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica : Journal of the Econometric Society*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Klein, G. (1999). Sources of power: How people make decisions. Cambridge, MA: MIT Press.

Kloosterman, N. A., Kosciessa, J. Q., Lindenberger, U., Fahrenfort, J. J., & Garrett, D. D. (2020). Boosts in brain signal variability track liberal shifts in decision bias. *eLife*, *9*, e54201. https://doi.org/10.7554/eLife.54201

Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management*, *30*, 15–29.

Lughofer, E. D. (2008). FLEXFIS: A robust incremental learning approach for evolving Takagi–Sugeno fuzzy models. *IEEE Transactions on Fuzzy Systems*, *16*(6), 1393–1410.

Makrehchi, M., & Kamel, M. S. (2011). An information theoretic approach to generating fuzzy hypercubes for IF-THEN classifiers. *Journal of Intelligent & Fuzzy Systems*, *22*(1), 33–52.

Norman, D. A. (2014). Some observations on mental models. In *Mental models* (pp. 7–14). Psychology Press.

Pickering, L., Cohen, K., & De Baets, B. (2025). A narrative review on the interpretability of fuzzy rule-based models from a modern interpretable machine learning perspective. *International Journal of Fuzzy Systems*, (pp. 1–20). https://doi.org/10.1007/s40815-025-02022-z

Plonka, L., & Mrozek, A. (1995). Rule-based stabilization of the inverted pendulum. *Computational Intelligence*, *11*(2), 348–356.

Rastogi, C., Leqi, L., Holstein, K., & Heidari, H. (2023). A taxonomy of human and ML strengths in decision-making to investigate human-ML complementarity. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *11*(1), 127–139.

Ryu, H. S., Ju, U., & Wallraven, C. (2021). Predicting decision-making in the future: Human versus machine. In *Asian conference on pattern recognition* (pp. 127–141). Springer.

Scherer, C., & Weiland, S. (2000). Linear matrix inequalities in control. Lecture Notes, Dutch Institute for Systems and Control. Available: http://www.cs.ele.tue.nl/sweiland/lmi.htm.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Sudár, A., Berki, B., & Horváth, I. (2025). Fuzzy model-based analysis of user feedback for product development insights. *Heliyon*, *11*(12), e43537. https://doi.org/10.1016/j.heliyon.2025.e43537

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge, MA: MIT Press.

Tanaka, K., & Wang, H. O. (2001). Fuzzy control systems design and analysis: A linear matrix inequality approach. John Wiley and Sons.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Vasudevan, V., & Ramakrishna, M. (2017). A hierarchical singular value decomposition algorithm for low rank matrices. arXiv:1710.02812.

Wu, T.-P., & Chen, S.-M. (1999). A new method for constructing membership functions and fuzzy rules from training examples. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *29*(1), 25–40.

Yam, Y., Baranyi, P., & Yang, C. T. (1999). Reduction of fuzzy rule base via singular value decomposition. *IEEE Transactions on Fuzzy Systems*, *7*(2), 120–132. https://doi.org/10.1109/91.755394

Zadeh, L. A. (1996). Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems*, *4*(2), 103–111.

Zapata, G. O. A., Kawakami, R., Galvao, H., & Yoneyama, T. (1999). Extracting fuzzy control rules from experimental human operator data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *29*(3), 398–406.

Zarandi, M. H. F., Turksen, I. B., & Rezaee, B. (2004). A systematic approach to fuzzy modeling for rule generation from numerical data. In *IEEE Annual meeting of the fuzzy information, processing Nafips'04* (pp. 768–773). IEEE (*Vol. 2*).

Zhang, W., Deng, Z., Wang, G., & Choi, K.-S. (2025). Fuzzy rule-based differentiable representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(1), 1–14.