When Noticing Is Not a Command: Negotiated Agency in AI-Driven Assistive Household Cleaning Robots

Toan Quoc Nguyen
toanquoc.nguyen@tu-dortmund.de
TU Dortmund University
Dortmund, Germany

Sven Mayer
sven.mayer@tu-dortmund.de
TU Dortmund University
Dortmund, Germany

Abstract

Household cleaning robots have the potential to support more independent living for individuals with motor impairments. However, most existing systems rely on explicit control interfaces or rigid, schedule-based operation, which are often inaccessible for these users and poorly aligned with the dynamic nature of domestic environments. Cleaning needs arise unpredictably, and both delayed intervention and intrusive robot behaviour can undermine safety, well-being, and user trust. Electroencephalography (EEG) offers a hands-free interaction modality that captures neural signals without requiring overt movement or speech, making it a promising alternative for accessible household robot interaction. Yet prior work in brain-computer interfaces has identified the Midas Touch problem, in which transient perceptual or cognitive appraisals are misinterpreted as intentional commands, leading to accidental or unwanted actions. This paper addresses this challenge by proposing a negotiated agency framework for EEG-based household cleaning robots. Grounded in human-centred and value-sensitive design, the framework employs AI-inferred EEG signals as probabilistic intent proposals rather than direct commands. By integrating AI-driven EEG inference with soft-intervention mechanisms for implicit confirmation or veto, the approach enables context-aware assistance while preserving human agency and autonomy.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Human-AI Alignment, Brain-Computer Interface, EEG, Household Cleaning Robot, Midas Touch, Accessibility, Human-AI Interaction

ACM Reference Format:
Toan Quoc Nguyen and Sven Mayer. 2026. When Noticing Is Not a Command: Negotiated Agency in AI-Driven Assistive Household Cleaning Robots. In *Proceedings of Workshop on Human-AI Interaction Alignment: Designing, Evaluating, and Evolving Value-Centered AI For Reciprocal Human-AI Features (CHI '26).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

CHI '26, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Household cleaning robots are increasingly deployed to support daily living in domestic environments, with particular promise as assistive tools for individuals with motor impairments [1, 2]. Tasks such as vacuuming, wiping, or responding to spills may appear routine, yet they pose substantial barriers to independent living for people with limited mobility and reduced dexterity. Autonomous cleaning robots have the potential to alleviate physical strain, enhance safety, and support dignity by enabling individuals to maintain clean living environments with reduced dependence on caregivers, while easing the burden of routine, physically demanding household cleaning tasks in a context-aware manner [3–6].

Despite these benefits, effective interaction with household cleaning robots remains a critical challenge. Most current systems rely on explicit interaction modalities, including physical buttons, mobile applications, or voice commands [1]. Such interfaces can be inaccessible for users with motor and/or speech impairments, particularly in everyday situations. Even seemingly lightweight confirmation actions may impose non-trivial physical or significant cognitive effort, limiting their suitability for possible efficient interaction.

Moreover, routine schedule-based cleaning overlooks situational context, while continuous, always-on operation can be intrusive, energy-inefficient, and disruptive to everyday life [7]. Cleaning needs often arise unpredictably, such as liquid spills caused by children or pets, food crumbs during meals or snacks, dust or debris carried in by foot traffic or wind through open windows, mud or water tracked indoors after rain, hygiene-critical contamination following accidental messes, to name but a few. However, constant robot patrolling may undermine privacy, rest, and a sense of calm, particularly when robot behaviour is perceived as overly present or socially inappropriate [8]. Conversely, delayed responses to genuine cleaning needs can introduce safety hazards, bacterial growth, and reduced indoor air quality, with disproportionate consequences for individuals with limited mobility or compromised immune systems. The challenge is therefore not simply to clean more often, but to clean at the *right time*, in ways that respect users' situational priorities and tolerance for interruption [7, 8].

Brain–computer interfaces (BCIs), and in particular electroencephalography (EEG), offer an alternative interaction channel that enables hands-free access to neural signals associated with perception and evaluation for controlling robots [2, 9–12]. EEG-based interaction does not require overt motor actions or speech and can operate alongside everyday activities, making it especially relevant for users with motor and/or speech impairments. When combined with artificial intelligence (AI), EEG has the potential to support implicit, background interaction in domestic environments. Nevertheless, the research in advancing an EEG-based AI-driven interface

for controlling domestic cleaning robots supporting differently-abled communities is still *under-explored*, leading to a research gap. Hence, this paper may ignite the field to promote development for these vulnerable groups, assisting their more independent life.

However, EEG-based AI-driven approaches that treat neural signals as direct control commands risk introducing the *Midas Touch* problem, in which transient perceptual or evaluative responses are misinterpreted as intentional commands, leading to accidental or unwanted system actions [13–17]. Noticing a spill does not necessarily imply a desire for immediate intervention; users may be resting, hosting guests, or prioritising other activities. Systems that respond automatically to every detected appraisal risk increase cognitive load, cause unwanted interruptions, and erode trust.

Recent work on human-AI alignment emphasises that alignment is not a one-time optimisation problem but a dynamic, bidirectional process that unfolds through interaction and adaptation over time [18, 19]. Beyond mis-specified objectives, misalignment often arises from fundamental differences in how humans and AI systems interpret, generalise, and act on signals in context, particularly under uncertainty [20]. From this perspective, effective alignment requires interaction-level mechanisms that enable humans to implicitly critique, endorse, or veto AI behaviour during use, especially in real-world settings where values, intentions, and tolerances for intervention are situational and cannot be fully specified or reliably inferred in advance [19, 21].

In this paper, based on Value Sensitive Design (VSD) [22–24], we argue that EEG-based interaction with household cleaning robots should be reframed as a process of *negotiated agency*. Rather than considering AI-classified EEG signals as deterministic commands, we propose interpreting them as probabilistic *intent proposals* that initiate lightweight, interaction-level negotiation. The robot responds to inferred relevance with soft, minimally intrusive pre-action cues, enabling users to implicitly confirm or veto proposed actions without physical movement or speech. We present a value-centred framework that integrates AI-driven EEG interpretation with interaction-level decision rules to preserve human agency while enabling accessible, context-aware assistance for independent living. In summary, this research aims to address the following research questions (RQs):

- **RQ1:** How do neural signals associated with *perceptual awareness* and *volitional endorsement* manifest during household cleaning scenarios, and how can AI models leverage these patterns to distinguish between them?
- **RQ2:** How do soft-intervention cues mediate users' implicit acceptance or veto of AI-proposed robot actions across different domestic and social contexts?
- **RQ3:** How can negotiated, AI-driven decision rules balance robot autonomy and human agency while reducing false activations in household cleaning interaction?

2 Related Work

2.1 EEG-Based Robot Interaction and Implicit Control

Prior work on EEG-based robot interaction with integrated AI models has largely relied on explicit command paradigms [9], including motor imagery [25] and steady-state visual evoked potentials (SSVEP) [26, 27], to enable discrete command selection. While these approaches demonstrate high performance in controlled laboratory environments, they impose substantial cognitive and attentional demands on users, requiring sustained focus, deliberate mental effort, and extensive calibration or training. Such requirements are difficult to maintain in everyday domestic settings, where attention is fragmented, interaction is intermittent, and users may experience fatigue, cognitive load, or motor impairments. Beyond EEG-specific constraints, prior work in robot learning has similarly shown that explicit human feedback is inherently limited in complex and dynamic task situations, as it requires deliberate evaluation, incurs additional decision time, and becomes unreliable in ambiguous cases where robot behaviour is neither clearly correct nor incorrect [28]. Consequently, explicit command- and feedback-based interaction paradigms face significant barriers to practical deployment beyond experimental contexts.

To address these limitations, more recent work has explored implicit neural signals as supervisory input for autonomous systems, aiming to reduce reliance on the deliberate generation of commands [28, 29]. However, most of these approaches treat neural responses as immediate triggers for system action, effectively collapsing *perceptual awareness* [30, 31], the user's noticing or appraisal of a situation, and *volitional endorsement* [32], the user's permission to act, into a single control signal. As a result, such systems remain susceptible to unintended activation and fail to account for the fundamental ambiguity between noticing a situation and intending to act, undermining safety, trust, and reliability in real-world human-robot interaction (HRI).

2.2 Human Agency and Accidental Activation in Hands-Free Interaction

Within human-robot interaction, prior work has highlighted the risks of accidental activation, over-automation, and miscalibrated trust in hands-free and autonomous systems, particularly when system behaviour is insufficiently aligned with user intent and situational context [33]. In assistive and ubiquitous computing settings, such unintended actions can be especially disruptive, as they increase cognitive load, generate confusion, and undermine user trust and acceptance [34]. Recent HRI research further demonstrates that user trust and acceptance are highly sensitive to how robotic behaviour aligns with human expectations, values, and perceived appropriateness, with misalignment leading to reduced trust and degraded interaction quality [35].

These concerns are further amplified in EEG-based interaction paradigms. EEG signals are inherently noisy, non-stationary, and highly sensitive to contextual and user-state variations, making reliable inference of intention fundamentally ambiguous [9]. When coupled with embodied robotic systems operating in dynamic environments, where unintended physical actions can carry safety and usability consequences [36], EEG-driven interfaces face a heightened risk of false activations and over-responsiveness. This reinforces the need for interaction designs that explicitly separate perception, evaluative judgment, and intentional control, to preserve user agency, ensure appropriate trust calibration, and reduce unintended system behaviour.

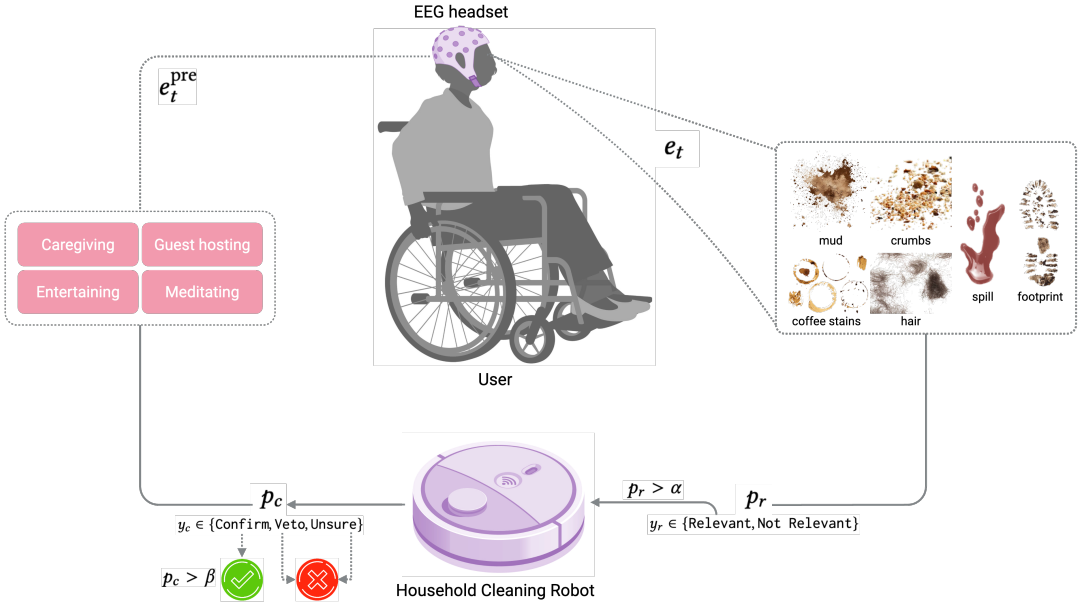

Figure 1: Overview of the proposed *Negotiated Agency Framework* for EEG-based household cleaning robots. During *stimulus observation*, neural signals (e_t) are recorded while the user perceives the environment (e.g., dry debris or liquid spills). A first-stage inference estimates *perceptual relevance* (p_r, $y_r \in \{\texttt{Relevant}, \texttt{Not Relevant}\}$), indicating whether the situation is appraised as potentially requiring cleaning. When relevance is inferred, the robot issues a *pre-action cue* that externalises its intent proposal without committing to action. Following this cue, EEG signals (e_t^{pre}) are used in a second-stage inference to estimate *volitional endorsement* (p_c, $y_c \in \{\texttt{Confirm}, \texttt{Veto}, \texttt{Unsure}\}$), reflecting implicit acceptance, rejection, or hesitation. Robot action is executed only when both relevance and confirmation are inferred; veto and hesitation default to inaction. Social and situational contexts (e.g., caregiving, guest hosting, meditation) modulate endorsement, ensuring context-aware, human-centred alignment under uncertainty.

VSD provides a principled framework for addressing these challenges by foregrounding human values such as autonomy, dignity, and freedom from interruption. However, many EEG-driven systems incorporate values only implicitly, embedding them at the level of model optimisation rather than at the level of interaction. Consequently, users are often left without effective means to negotiate or contest system behaviour during use.

2.3 Bidirectional Human-AI Alignment and Interaction-Level Negotiation

Recent work on human-AI alignment increasingly characterises alignment not as a one-time optimisation problem, but as an ongoing, reciprocal process that unfolds through interaction, feedback, and adaptation over time [19, 21]. From this perspective, misalignment is not solely attributable to misspecified objectives or insufficient training data, but frequently emerges during deployment, when AI systems encounter novel contexts, distributional shifts, and underspecified or evolving human preferences. In particular, prior work has shown that systematic differences in how humans and AI systems interpret and generalise from signals and experiences can lead to divergence between perceived relevance and intended action, even when high-level goals appear to be shared [20].

Despite these insights, much of the existing alignment literature continues to prioritise model-level mechanisms, such as preference learning, reward shaping, or post hoc evaluation, while paying comparatively less attention to interaction-level processes that allow humans to shape system behaviour *in situ* throughout use [21]. Even within human-AI and HRI research, alignment is often implicitly assumed to occur through explicit feedback, corrective input, or deliberate command signals. However, prior work in interactive robot learning cautions that such assumptions rely on treating humans as reliable oracles, despite cognitive biases, ambiguity, and inconsistency in human judgment and behaviour [37]. As a result, explicit command- or feedback-based alignment mechanisms remain brittle in complex, real-world settings, and offer limited support for expressing uncertainty, deferring action, or contesting system decisions without disrupting ongoing activity.

Our work addresses this gap by operationalising negotiated agency in the context of EEG-based household robotics. Rather than using neural signals as direct or deterministic control commands, we embed AI-driven EEG mechanisms within a two-stage interaction loop that explicitly separates perceptual appraisal from action execution. This design enables context-aware alignment by allowing AI systems to surface tentative action proposals while preserving the human's ability to implicitly endorse or veto them

during use. In doing so, our approach advances bidirectional alignment research by situating alignment at the level of interaction, treating neural signals as conversational and probabilistic input rather than deterministic control, and grounding alignment mechanisms within the temporal, contextual, and value-laden realities of domestic environments [19–21, 37].

3 Method

This work proposes a *Negotiated Agency Framework* that operationalises bidirectional human-AI alignment for household cleaning robots (See Figure 1). Grounded in the *VSD tripartite methodology* [22–24], the framework integrates three interdependent investigations: (i) a *conceptual investigation* of stakeholder value tensions in domestic autonomy, (ii) an *empirical investigation* using a Wizard-of-Oz (WoZ) paradigm [38–40] to examine lived value negotiations, and (iii) a *technical investigation* that formalises AI-mediated EEG inference as a value-representative control logic.

By employing this tripartite structure, we ensure that the technical mechanisms, specifically the thresholds for robot intervention, are not arbitrary but are grounded in human values and validated through interaction-level evidence. Central to this framework is the ontological separation of *perceptual awareness*, *intent proposal*, and *action execution*. EEG is treated not as a direct control modality, but as a probabilistic indicator of human appraisal, allowing the system to preserve human agency under uncertainty.

3.1 Conceptual Foundations: Stakeholder Values and Situated Autonomy

The conceptual investigation establishes the normative foundations of the framework by identifying the values at stake and the stakeholders they affect. Following VSD principles, we distinguish between *direct stakeholders*, such as the primary users with motor and/or speech impairments, and *indirect stakeholders*, including co-habitants, guests, or caregivers who are impacted by the robot's presence in the domestic sphere.

We identify a fundamental *value tension* between *Cleanliness* and *Calm and Non-Intrusion*:

- *Cleanliness* encompasses hygiene, safety, and physical well-being. For direct stakeholders with physical impairments, this value is tied to *functional independence* because they may be unable to address hazards, such as liquid spills, without robotic assistance.
- *Calm and Non-Intrusion* represents dignity, privacy, and freedom from interruption. This value ensures the home remains a space for rest and social interaction, free from over-vigilant automation that may feel patronising or intrusive to both direct and indirect stakeholders.

While traditional robotic systems often treat cleanliness or efficiency as a singular optimisation objective, our framework treats alignment as a *situated and negotiable process*. We argue that the relative priority of these values is dynamic and contingent on environmental urgency (e.g., a slip hazard versus dust) and social context (e.g., being alone versus hosting guests).

This conceptual framing provides the requirement for our technical design. The system must not act on perceptual triggers alone. Instead, it must engage in a negotiation where the robot's agency is contingent on human appraisal. By formalising this tension, the framework avoids the *Midas Touch problem*, where every perceived stimulus is misinterpreted as a command. Instead, it privileges the user's right to remain undisturbed unless a clear endorsement is inferred.

3.2 Empirical Investigation: Wizard-of-Oz Study

To examine negotiated agency at the level of lived interaction, we propose a WoZ study [38–40] that simulates an AI-driven, EEG-informed household cleaning robot while retaining experimental control over decision points (See Figure 1). The WoZ methodology can be employed to isolate interaction dynamics under uncertainty, without conflating user experience with the development of real-time EEG decoding.

The primary objective of this study is not to evaluate EEG classification performance, but to understand how users perceive, interpret, and respond to robot-initiated negotiation cues when intent is ambiguous. While fully automated EEG-based systems would ultimately use neural signals for online inference, such systems need to handle unconstrained domestic environments. WoZ enables the robot to appear autonomous while allowing systematic exploration of interaction timing, cue interpretation, and decision outcomes under controlled conditions.

Experimental Design. The study follows a within-subject design. Participants observe a sequence of household cleaning scenarios while wearing an EEG headset. Scenarios are systematically varied along three orthogonal dimensions:

- **Environmental condition:** liquid spill, dry debris (e.g., crumbs or dust), mixed debris, or no cleaning required.
- **Urgency level:** high (e.g., slip hazard or spreading liquid) versus low (cosmetic or peripheral dirt).
- **Social context:** alone and relaxed, engaged in focused activity, or hosting guests.

These dimensions are derived directly from the conceptual value analysis and enable examination of how situational context modulates perceived relevance, tolerance for interruption, and willingness to accept robot intervention. Identical physical stimuli may appear across different contextual conditions, ensuring that differences in response arise from value-based appraisal rather than stimulus salience alone.

Interaction Flow and Data Classes. Each trial proceeds according to the following algorithmic sequence, which mirrors the intended autonomous system while remaining under WoZ control. At each stage, interaction data are associated with clearly defined semantic classes.

(1) **Stimulus Observation (Perceptual Appraisal).** The participant observes the environment while the EEG is recorded. The robot remains idle, and no cues are presented. EEG segments recorded during this phase are associated with a *perceptual relevance* label:

$$y_r \in \{\texttt{Relevant}, \texttt{Not Relevant}\},$$

indicating whether the participant appraised the scene as potentially requiring cleaning. These labels are derived from scenario design and post-trial reports.

(2) **Pre-Action Cue (WoZ-Controlled Intent Proposal).** When a scenario is designed to plausibly elicit cleaning relevance (i.e., y_r = `Relevant`), the wizard triggers a *pre-action cue*. The cue is intentionally subtle and may be delivered through a brief auditory signal (e.g., a soft chime or tonal pulse), a minimal visual change (e.g., a light indicator or orientation shift), or a gentle haptic feedback. It serves to externalise the robot's internal deliberation without demanding attention or action. Importantly, this cue constitutes a *proposal*, not a command, and does not commit the robot to action.

(3) **Implicit Human Response (Volitional Appraisal).** EEG activity following the pre-action cue is recorded. These EEG segments are associated with a second semantic label capturing the participant's volitional response to the proposal:

$$y_c \in \{\texttt{Confirm}, \texttt{Veto}, \texttt{Unsure}\}.$$

Here, `Confirm` indicates endorsement of robot action, `Veto` indicates rejection or disengagement, and `Unsure` reflects hesitation or unresolved intent. During the WoZ study, these labels are not inferred online but are assigned offline using post-trial self-reports.

(4) **Action Resolution (WoZ-Controlled).** Depending on the experimental condition, the wizard either initiates the cleaning action or withholds action. Inaction is treated as the default outcome unless confirmation is explicitly designed into the condition. Both `Veto` and `Unsure` are mapped to inaction, reflecting a context-aware alignment strategy. To the participant, the robot appears to resolve the decision autonomously.

(5) **Post-Trial Self-Report (Ground Truth).** After each trial, participants report whether they would have wanted the robot to act (`Confirm`), preferred it not to act (`Veto`), or felt uncertain (`Unsure`). Confidence ratings may additionally be collected. These self-reports serve as ground truth for validating interaction outcomes and for labelling EEG data corresponding to both perceptual relevance (y_r) and volitional response (y_c).

Confirmation, Veto, and Hesitation Conditions. Stimuli are constructed such that identical physical dirt or debris may elicit different volitional classes depending on context. Trials intended to elicit `Confirm` responses pair cleaning stimuli with contexts in which intervention is typically desirable and low-cost (e.g., being alone, minimal cognitive engagement, potential safety risk). Trials intended to elicit `Veto` responses pair identical stimuli with contexts where intervention would plausibly violate values of peace, dignity, or social appropriateness (e.g., hosting guests, resting, meditating). A third class of trials is designed to induce `Unsure` responses, reflecting situations in which participants may hesitate or feel ambivalent about whether cleaning should occur.

Critically, the absence of confirmation following a pre-action cue is treated as a veto. This design choice reflects both ethical and practical considerations: under uncertainty, preserving peace, privacy, and freedom from interruption is preferable to risking unwanted intervention. Hesitation is therefore treated as a meaningful interaction outcome rather than as noise.

Design Rationale. Two principles govern this WoZ design. First, perceptual awareness is explicitly separated from action execution: noticing dirt is insufficient to trigger cleaning. Second, agency is preserved through context-aware decision logic that privileges inaction in the absence of clear endorsement. By embedding negotiation at the interaction level, the study enables examination of how users experience AI behaviour that remains contingent on human judgement rather than deterministic automation.

The WoZ study thus provides empirical grounding for the negotiated agency framework. Observed class distributions and user responses inform the calibration of relevance and confirmation thresholds, veto logic, and feedback modalities in the algorithmic model, ensuring that subsequent automation is shaped by empirically observed human values rather than purely optimisation-driven assumptions.

3.3 Technical Investigation: Negotiated Agency Through Two-Stage Intent Inference

This component specifies how negotiated agency is operationalised as an explicit, implementable decision model (See Figure 1). At the algorithmic level, alignment is realised through a two-stage AI-mediated mechanism that deliberately separates *perceptual appraisal* from *volitional endorsement*. This separation ensures that noticing a potential cleaning situation is never conflated with an intention to act.

Stage 1: Perceptual Relevance Inference (Intent Proposal). Let e_t denote an EEG segment recorded while the user observes the environment at time t. A first classifier $f(\cdot)$ maps features extracted from e_t to a probabilistic estimate of perceived cleaning relevance:

$$p_r = \sigma(f(e_t)),$$

where $\sigma(\cdot)$ denotes the sigmoid function. This output corresponds to a binary perceptual appraisal:

$$y_r \in \{\texttt{Relevant}, \texttt{Not Relevant}\}.$$

Importantly, p_r represents an *intent proposal* rather than a command. A high relevance probability indicates that the user has noticed or appraised a potential cleaning situation, but does not imply any desire for immediate robot action.

When p_r exceeds a relevance threshold α, the system is authorised only to externalise its deliberation by issuing a pre-action cue. No action is executed at this stage. If $p_r \leq \alpha$, the robot remains idle.

Stage 2: Volitional Endorsement Inference (Action Gating). Following the pre-action cue, a second EEG segment e_t^{pre} is recorded. A second classifier $g(\cdot)$ estimates the user's volitional response to the proposal:

$$p_c = g(e_t^{\text{pre}}),$$

corresponding to a multi-class endorsement label:

$$y_c \in \{\texttt{Confirm}, \texttt{Veto}, \texttt{Unsure}\}.$$

Here, `Confirm` indicates implicit endorsement of robot action, `Veto` indicates rejection or disengagement, and `Unsure` reflects unresolved or hesitant intent. Critically, only `Confirm` authorises execution; both `Veto` and `Unsure` map to inaction.

Robot action is therefore executed if and only if:

$$p_r > \alpha \quad \text{and} \quad p_c > \beta,$$

where β is a confirmation threshold associated with the `Confirm` class. In all other cases, the system defaults to inaction.

Context-aware Alignment Strategy. This two-stage gating mechanism is intentionally context-aware. It ensures that perceptual awareness alone cannot trigger cleaning and that the absence of clear endorsement is treated as a veto. By privileging inaction under uncertainty, the model directly addresses the Midas Touch problem, preventing transient cognitive appraisals or ambiguous neural responses from being misinterpreted as intentional commands.

Relation to the Wizard-of-Oz Study. During the WoZ study, both classifiers are conceptual rather than operational: pre-action cues and action outcomes are controlled by the experimenter, while EEG signals are recorded for offline analysis and labelling. Post-trial self-reports provide ground truth for both perceptual relevance (y_r) and volitional endorsement (y_c). This design allows the interaction logic and class structure of the algorithm to be validated independently of real-time decoding performance.

Adaptation and Personalisation. The thresholds α and β are value-sensitive parameters reflecting priorities identified in the conceptual analysis and observed in the WoZ study. In future deployments, these thresholds may be adapted over time to support personalisation and long-term bidirectional alignment, allowing the system to learn individual preferences while maintaining safety, trust, and low interaction burden.

Together with the conceptual and empirical components, this algorithmic model may complete a coherent methodological pipeline: values motivate interaction design, interaction design constrains algorithmic authority, and algorithmic mechanisms operationalise negotiated agency in a form suitable for real-world deployment.

4 Discussion and Conclusion

This work introduces a *Negotiated Agency Framework* for EEG-based household cleaning robots that advances human-robot interaction beyond task execution toward interaction-level alignment. Rather than addressing how robots optimise or execute cleaning actions, the proposed method focuses on how assistive robots should *decide when to act* in domestic environments characterised by uncertainty, social context, and competing human values. In response to **RQ1**, the framework establishes a principled distinction between neural signals associated with *perceptual awareness* and those reflecting *volitional endorsement*, reconceptualising EEG not as a direct control channel but as a probabilistic indicator of human appraisal embedded within a context-aware, two-stage decision logic. This separation enables AI models to distinguish between noticing and permission, preventing perceptual signals from being prematurely interpreted as commands and addressing the *Midas Touch problem*.

At the interaction level, and addressing **RQ2**, the framework operationalises negotiation as an explicit and integral design mechanism. Instead of relying on explicit commands or continuous automation, the robot surfaces tentative action proposals through soft, minimally intrusive cues and defers execution unless endorsement is inferred. This interaction design enables users to implicitly accept, veto, or hesitate in response to AI-proposed actions across different domestic and social contexts, without physical movement or speech. By treating hesitation and non-response as meaningful interaction outcomes rather than noise, the framework reflects the situated and value-laden nature of domestic interaction.

Finally, in response to **RQ3**, the proposed method embeds context-aware, AI-driven decision rules that balance robot autonomy with human agency while reducing false activations in household cleaning interaction. Robot action is authorised only when both relevance and endorsement are inferred; otherwise, the system defaults to inaction. By privileging non-intervention under uncertainty, the framework preserves users' autonomy, dignity, and right to remain undisturbed, while still enabling timely assistance when endorsement is present.

Beyond the specific application of household cleaning, this work contributes a generalisable approach to bidirectional human-AI alignment in hands-free and neuroadaptive systems. By embedding value-driven constraints directly into interaction logic and algorithmic authority, the framework shifts alignment from a model-level optimisation problem to an interaction-level process that unfolds during use. This perspective is particularly relevant for assistive technologies intended for individuals with motor and/or speech impairments, where accessibility must be balanced against the risks of over-automation and loss of agency.

Limitations and Future Work. The proposed method is intentionally context-aware and interaction-focused. While this design prioritises autonomy preservation and safety, it may delay beneficial intervention in borderline situations where user appraisal remains ambiguous. Future work should therefore investigate adaptive strategies that retain context-aware defaults while enabling gradual personalisation over time.

On top of that, the present framework centres on a core value tension between cleanliness and calm. Extending the method to incorporate a broader set of stakeholders, values, and long-term household dynamics represents an important direction for further research. Moreover, the development of dedicated evaluation metrics will be essential to assess not only AI model performance within the technical investigation but also user experience, trust, and perceived agency during interaction. Beyond standard classification metrics (e.g., accuracy, false positive rate, and latency), future evaluations should incorporate interaction-level measures such as perceived control, cognitive load, trust calibration, and the frequency of unintended or missed actions. Longitudinal and scenario-based user studies will be particularly important to capture how users adapt to the negotiation mechanism over time and how the system influences autonomy, comfort, and acceptance in everyday environments.

Finally, although the framework is articulated independently of specific EEG decoding techniques, its effectiveness depends on the practical separability of neural correlates associated with perceptual awareness and volitional endorsement. In real-world settings, these signals are likely to be noisy, partially overlapping, and subject to substantial inter-subject and intra-subject variability. This introduces a fundamental trade-off between sensitivity (detecting true endorsement) and specificity (avoiding false activations), particularly under low signal-to-noise conditions. Future work should therefore explicitly investigate the feasibility of this distinction through empirical studies, including analyses of signal separability, temporal dynamics, and robustness across users. Hybrid approaches

that combine EEG with complementary modalities (e.g., contextual sensing or behavioural cues) may further improve reliability while preserving the hands-free interaction paradigm.

References

[1] Yoshiaki Shiokawa, Winnie Chen, Aditya Shekhar Nittala, Jason Alexander, and Adwait Sharma. Beyond vacuuming: How can we exploit domestic robots' idle time? In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25')*, pages 1–17, 2025.

[2] Asif Iqbal, Ashok Kumar Suhag, Neeraj Kumar, and Arpit Bhardwaj. Use of BCI Systems in the Analysis of EEG Signals for Motor and Speech Imagery Task: A SLR. *ACM Computing Surveys*, 58(2):1–25, 2025.

[3] Rajesh Kannan Megalingam, Shree Rajesh Raagul Vadivel, Sai Smaran Kotaprolu, Bagathi Nithul, Devisetty Vijay Kumar, and Gaurav Rudravaram. Cleaning robots: A review of sensor technologies and intelligent control strategies for cleaning. *Journal of Field Robotics*, 2025.

[4] Gebrezabher Niguse Hailu, Haftu Berhe Gebru, Gebrezgiher Gidey Hagos, Abrha Hailay Weldemariam, Degena Bahrey Tadesse, and Guesh Mebrahtom. The role of family caregivers in supporting older adults in africa: systematic review. *BMC geriatrics*, 25(1):491, 2025.

[5] Shujing Suo, Jiawei Jiao, Yun Li, Jinjin Gu, Shitong Guo, Chang Liu, Shiguang Wang, Panpan Wang, and Peng Wang. Development and validation of an index system for caregiving capacity of family caregivers in caring for older adults with disabilities at home: a delphi and analytic hierarchy process. *BMC nursing*, 24(1):818, 2025.

[6] Julia Smith, Alice Mŭrage, Kayli Jamieson, and Kaylee A Byers. "not having the energy to even live": A feminist disability perspective on long covid and caregiving. *Health & Social Care in the Community*, 2025(1):8893161, 2025.

[7] Morten Hertzum. Inferior, yet transformative: the user experience with robotic vacuum cleaners. *Interacting with Computers*, 36(1):16–29, 2024.

[8] Bram Hendriks, Bernt Meerbeek, Stella Boess, Steffen Pauws, and Marieke Sonneveld. Robot vacuum cleaner personality and behavior. *International Journal of Social Robotics*, 3(2):187–195, 2011.

[9] Yuchong Zhang, Nona Rajabi, Farzaneh Taleb, Andrii Matviienko, Yong Ma, Mårten Björkman, and Danica Kragic. Mind meets robots: a review of EEG-based brain-robot interaction systems. *International Journal of Human–Computer Interaction*, pages 1–32, 2025.

[10] Yidan Ding, Chalisa Udompanyawit, Yisha Zhang, and Bin He. Eeg-based brain-computer interface enables real-time robotic hand control at individual finger level. *Nature Communications*, 16(1):1–20, 2025.

[11] Ajay Joshi, Lynne Baillie, and Carl Bettosi. A robot-administered ICU confusion assessment with Brain-Computer Interface control. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24')*, pages 583–587, 2024.

[12] Liu Shilong, Chaorui Tong, Zelu Liu, Xiangxian Li, Yawen Zheng, Chao Zhou, Juan Liu, and Yulong Bian. Assessing Dynamic Flow Experience from EEG Signals: A Processing-based Approach. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25')*, pages 1–19, 2025.

[13] G S Rajshekar Reddy, Michael J Proulx, Leanne Hirshfield, and Anthony Ries. Towards an eye-brain-computer interface: Combining gaze with the stimulus-preceding negativity for target selections in xr. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24')*, 2024.

[14] Tan Gemicioglu, R. Michael Winters, Yu-Te Wang, Thomas M. Gable, Ann Paradiso, and Ivan J. Tashev. Gaze & tongue: A subtle, hands-free interaction for head-worn devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*, 2023.

[15] Jalynn Blu Nicoly. Towards Seamless Interaction: Neuroadaptive Virtual Reality Interfaces for Target Selection. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI '25')*, page 745–748, 2025.

[16] Haruaki Fukuda, Masahiro Shiomi, Kayako Nakagawa, and Kazuhiro Ueda. 'Midas touch' in human-robot interaction: evidence from event-related potentials during the ultimatum game. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '12)*, page 131–132, 2012.

[17] Robin Schweigert, Valentin Schwind, and Sven Mayer. Eyepointing: A gaze-based selection technique. In *Proceedings of Mensch und Computer 2019*, pages 719–723. 2019.

[18] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Michael Xieyang Liu, Andrés Monroy-Hernández, Tongshuang Wu, Diyi Yang, Yun Huang, Tanushree Mitra, Yang Li, and Marti Hearst. Bidirectional Human-AI Alignment: Emerging Challenges and Opportunities. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*.

[19] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. Towards bidirectional Human-AI alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2406:1–56, 2024.

[20] Filip Ilievski, Barbara Hammer, Frank van Harmelen, Benjamin Paassen, Sascha Saralajew, Ute Schmid, Michael Biehl, Marianna Bolognesi, Xin Luna Dong, Kiril Gashteovski, et al. Aligning generalization between humans and machines. *Nature Machine Intelligence*, 7(9):1378–1389, 2025.

[21] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, et al. Ai alignment: A contemporary survey. *ACM Computing Surveys*, 58(5):1–38, 2025.

[22] Malak Sadek, Marios Constantinides, Daniele Quercia, and Celine Mougenot. Guidelines for integrating value sensitive design in responsible ai toolkits. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, pages 1–20, 2024.

[23] Theresa Schmiedel, Vivienne Jia Zhong, and Friederike Eyssel. Towards a wave approach for value sensitive design in social robotics. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23')*, pages 592–596, 2023.

[24] Nina Paukert, Carla Schurtenberger, Vivienne Jia Zhong, Janine Jäger, and Theresa Schmidel. Designing robots with values in mind: The role of colors in value-sensitive hri. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI '25')*, pages 1539–1543. IEEE, 2025.

[25] Arnau Dillen, Mohsen Omidi, Fakhreddine Ghaffari, Bram Vanderborght, Bart Roelands, Olivier Romain, Ann Nowé, and Kevin De Pauw. A shared robot control system combining Augmented Reality and motor imagery Brain–Computer Interfaces with Eye Tracking. *Journal of Neural Engineering*, 21(5):056028, oct 2024.

[26] Tianyi Yan, Zhiyuan Ming, Yilun Huang, Ziyu Liu, Qiming Chen, Deyu Zhang, Mengzhen Liu, Dingjie Suo, Jian Zhang, and Siyu Liu. Enhanced Brain-Controlled Mobile Robot Based on SE-VEP Paradigm With Single Stimulus. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2025.

[27] Lei Shao, Longyu Zhang, Abdelkader Nasreddine Belkacem, Yiming Zhang, Xiaoqi Chen, Ji Li, and Hongli Liu. EEG-Controlled Wall-Crawling Cleaning Robot Using SSVEP-Based Brain-Computer Interface. *Journal of Healthcare Engineering*, (1), 2020.

[28] Su Kyoung Kim, Elsa Andrea Kirchner, and Frank Kirchner. Flexible online adaptation of learning strategy using EEG-based reinforcement signals in real-world robotic applications. In *2020 IEEE International Conference on Robotics and Automation (ICRA '20')*, 2020.

[29] Alessandra Fava, Valeria Villani, and Lorenzo Sabattini. Error-related potentials in EEG signals: feature-based detection for human-robot interaction. *Scientific Reports*, 15(1):35183, 2025.

[30] Lynn C Robertson. Binding, spatial attention and perceptual awareness. *Nature Reviews Neuroscience*, 4(2):93–102, 2003.

[31] Michael A Cohen, Jonathan Keefe, and Timothy F Brady. Perceptual awareness occurs along a graded continuum: No evidence of all-or-none failures in continuous reproduction tasks. *Psychological Science*, 34(9):1033–1047, 2023.

[32] Ryan R Posh, Jonathan A Tittle, David J Kelly, James P Schmiedeler, and Patrick M Wensing. Hybrid volitional control of a robotic transtibial prosthesis using a phase variable impedance controller. In *2024 IEEE International Conference on Robotics and Automation (ICRA '24)*, pages 4555–4561. IEEE, 2024.

[33] Stephen L Dorton and Jeff C Stanley. Minding the gap: Tools for trust engineering of artificial intelligence. *Ergonomics in Design*, 33(3):142–147, 2025.

[34] Aakash Yadav, Sarah K Hopko, Prabhakar R Pagilla, and Ranjana K Mehta. Enhancing trust examinations with neural measures during human–robot collaboration under cognitive fatigue. *ACM Transactions on Human-Robot Interaction*, 15(2):1–27, 2025.

[35] Mona Mareen Kegel, Ruth Maria Stock-Homburg, and Tuure Tuunanen. Service robot morality in customer–robot interactions: A mixed-methods study. *ACM Transactions on Human-Robot Interaction*, 14(2), 2025.

[36] Robin Jeanne Kirschner, Kübra Karacan, Alessandro Melone, and Sami Haddadin. Categorizing robots by performance fitness into the tree of robots. *Nature Machine Intelligence*, 7(3):459–470, Feb 2025.

[37] Matthew Gombolay. Human-robot alignment through interactivity and interpretability: Don't assume a "spherical human". In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI '24')*, pages 8523–8528, 2024.

[38] Stephanie Kim, Jacy Reese Anthis, and Sarah Sebo. A taxonomy of robot autonomy for Human-Robot interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24')*, pages 381–393, 2024.

[39] Josh Bhagat Smith, Vivek Mallampati, Prakash Baskaran, Mark-Robin Giolando, and Julie A Adams. Design principles for building robust human-robot interaction machine learning models. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*, pages 247–251, 2024.

[40] Laurel D Riek. Wizard of Oz Studies in HRI: a systematic review and new reporting guidelines. *Journal of Human-robot Interaction*, 1(1):119–136, 2012.