Balancing Ethical Persuasion and Explanation Transparency in XAI: A Formal Framework for Adaptive User Interactions

Sukriti Bhattacharya[1 [0000−0002−2439−7546]], Rachele Carli[2 [0000−0002−8689−285X]], Igor Tchappi[3 [0000−0001−5437−1817]], Amro Najjar[1 [0000−0001−7784−6176]], and Davide Calvaresi[4 [0000−0001−9816−7439]]

[1] Luxembourg Institute of Science and Technology Maison de l'innovation 5, avenue des Hauts-Fourneaux, L-4362 Luxembourg
{sukriti.bhattacharya, amro.najjar}list.lu
[2] Responsible AI group, , , SE-901 87 Umeå, Sweden
rachele.carli@umu.se
[3] FINATRAX, SnT, University of Luxembourg, Esch-sur-Alzette, Luxembourg
igor.tchappi@uni.lu
[4] University of Applied Sciences Western Switzerland, Switzerland
davide.calvaresi@hevs.ch

Abstract. Explainable AI (XAI) systems must balance effectiveness with ethical constraints to prevent manipulative user interactions. Current frameworks often neglect how anthropomorphic design features and persuasive explanations may exploit cognitive vulnerabilities, risking unintended manipulation. We propose a formal framework for XAI, that dynamically balances ethical persuasion and explanation transparency. The framework integrates three core components: (1) a user vulnerability model quantifying susceptibility to manipulation, (2) a clarity-transparency trade-off mechanism adapting explanations to cognitive and contextual factors, and (3) time-decayed ethical bounds to prevent long-term manipulation. We validate the framework through a health coaching case study, demonstrating its ability to adjust persuasion strength based on user profiles and ethical constraints. Results confirm the framework's effectiveness in maintaining transparency, without compromising utility.

Keywords: XAI · responsible AI · anthropomorphization · persuasion · formalization.

1 Introduction

Explainable AI (XAI) has become indispensable for fostering trust and enabling critical evaluation of AI-driven decisions [25, 30]. By providing transparent and understandable explanations for AI decisions, XAI aims to build user trust and facilitate critical evaluation of AI recommendations [33, 12, 32, 42]. Therefore,

XAI models are often designed in such a way as to instill familiarity and reliability in the user, thus predisposing the latter to follow the instructions received. This is possible trying to create an understandable, familiar, and enjoyable interaction [9]. While the utility of such design features to ensure the operability, acceptance, and efficiency of the explanations is evident, it should be noted that some users may be particularly susceptible to interactions that emulate those with other trusted humans. This may induce or favor the creation of manipulative dynamics, in which the person is induced to behave in a way that they would not have otherwise [10], or to accept explanations that are intended to convince rather than to fill information or cognitive gaps [9]. In light of this, the challenge is finding the right balance between XAI's utility and users' protection from manipulative drift. More specifically, these explanations must be effective, but not manipulative.

While existing frameworks prioritize transparency, they often neglect the ethical implications of specific AI design features, particularly how explanations might exploit user vulnerability [16, 15]. This gap raises a critical question: *How can XAI models be effective while adhering to ethical constraints and avoiding manipulative drifts?*

We address this challenge by introducing a formal framework that dynamically balances persuasion strength with user-specific factors such as cognitive ability, vulnerability, and explanation fidelity [22, 11, 41, 26, 2, 18, 44]. Our approach integrates three core components: (a) a mathematical model quantifying user vulnerability through susceptibility, dependency, and autonomy metrics, (b) a clarity-transparency trade-off that adjusts explanations based on cognitive and contextual factors, and (c) time-decayed ethical bounds to prevent long-term manipulation. By grounding these mechanisms in formal methods [6, 43, 3, 29, 28], we ensure rigorous control over XAI persuasion. The following sections formalize these concepts, demonstrate their application in a health coaching scenario, and validate their ethical robustness.

Therefore, the paper is structured as follows: Section 2 analyzes the ethical risks of persuasive XAI, emphasizing the fine line between design for anthropomorphization and manipulation. Section 3 formalizes the framework's components, including user vulnerability metrics and the persuasion impact function. Section 4 details parameter estimation methods for real-world implementation. Section 5 outlines testable hypotheses guiding ethical XAI design. Section 6 validates the framework via a health coaching scenario, while Section 7 evaluates hypothesis compliance. Finally, Section 8 discusses implications and future directions.

2 Persuasion in XAI

XAI systems have the capacity to influence users' choices, thereby assisting them in adhering more consciously to the objective they set at the beginning of their interaction with an AI agent [36]. In this sense, XAI can be a tool to strengthen the persuasive power of the system with which one is interacting. Therefore, it is

essential that XAI provides instruments to allow the subject to assess whether they are acting in accordance with their goal or whether they wish to change it [35]. At the same time, XAI must not impose a certain outcome that has not been accepted *ex ante* by the user. That is to guarantee individuals' agency and self-determination over the decisions made, at each stage of the interaction [17].

To illustrate this point, consider the case of a behavior change recommendation system. The provision of effective recommendations has been demonstrated to enhance an individual's adherence to the habit-change program to which they have consented, thereby facilitating the initiation of system utilization. The explanation may in fact increase awareness of why one action is preferred over another, how this preference was selected, and what the actual consequences of such an attitude will be [40, 39]. A similar phenomenon may be observed in the context of AI systems designed for healthcare purposes, where explanations have been shown to positively influence therapeutic adherence [1]. For this purpose, it is critical that the user has a positive and open attitude towards the explanation and is able to trust it.

In order for this dynamic to be effective, studies in cognitive and behavioral sciences have demonstrated the relevance of certain design gimmicks, which make human-AI interaction as close as possible to the human relational experience [5]. Thus, it proves essential to equip XAI models with features that elicit the areas of the brain deputed to provoke feelings of trust, familiarity, and desirability towards the system [8].

2.1 Anthropomorphisation to Support Persuasive Purposes

The design gimmicks that contribute to making interaction linear, effective, and efficient are usually those capable of evoking a dimension of sociability and human-like interaction. In this regard, AI systems aimed at influencing users' actions and choices are often programmed to communicate not only with natural language [4], but also reproducing expressions usually used by a human being who cares about the interlocutor [19]. The emulation of feelings of attention, care, and interest is often made explicit — as in the case of expressions such as 'I know you so well', or 'I am here for you'. In other cases, we can find the use of emojis, user interfaces depicting virtual agents representing professional figures, or colors and fonts that target the areas of the brain deputed to retain information or, on the contrary, to lower cognitive defenses and act more on impulse [21]. These represent just a few examples of what has been defined in AI as anthropomorphisation. This expression identifies the set of design features implemented in order to elicit a response in the user in terms of attachment, reliance, and interest in the interaction with an artificial agent [7].

XAI models are designed following a similar logic [8], as their main purpose is to be engaging and comprehensible to a potentially wide range of non-specialized individuals. Hence, the use of natural language, as familiar as possible to the person involved in the interaction, but also of user interfaces that are appealing and that somehow convey reliability and trustworthiness with respect to the explanations provided.

These design features have a purely functional purpose, but act on people's subconscious sphere, thus escaping the scrutiny of critical thinking and conscious evaluation [14]. Therefore, they can have negative and potentially harmful side effects, along with utilitarian purposes. Among those, the one that could prove most problematic is the manipulative drift.

2.2 Manipulation as the Potential Dark Side of Persuasive Techniques

One of the most insidious negative side effects of any persuasive practice is manipulation. The main reason for this is that the distinction between these two practices is often very subtle.

Especially at the legal level, there have been many attempts to define what is meant by manipulative practices, in order to set clear boundaries between legitimate and illegitimate influence. However, the debate in doctrine and jurisprudence is not only still open, but has been made even more pressing by the advent of highly interactive AI technologies. Trying to draw a common denominator between the various theorisations of manipulation, one would have to refer to a type of influence that is difficult to resist [24], not because it is convincing or forcibly imposed, but because it appeals only and exclusively to cognitive-emotional structures over which individuals have no possibility of control or timely control [27, 24]. Take, for instances, the case of songs replayed in shopping centers as they possess frequencies that target the areas of the brain that regulate the impulse to buy, eliciting them [20]. Another example is that of images reproduced on the screen at such a speed that they do not allow conscious perception, yet target specific areas of the brain, inducing consequential behavior [34, 23].

Under such an influence, a subject is induced to do something that they would not have done without the influence itself, and that advantage in some ways the manipulator [13]. The latter regardless of the fact that it does not harm the manipulated person.

2.3 The Potential Manipulative Drifts of Designing for Anthropomorphisation

In the context of highly interactive AI systems — such as chatbots or recommender systems, for example — and the corresponding XAI models, the importance of design features capable of eliciting a psycho-emotional response in the user, thus orienting them towards collaboration rather than rejection of the technology, has been emphasized (Section 1). Such a design acts on the adaptive cognitive and psychological systems that have characterized the evolution of the human brain and its experience of the world outside the subject [31, 38]. These mechanisms are inherently shared across individuals, though their intensity and occurrence may vary. Moreover, they are not fully aware and, even if explained to people, they will remain not fully controllable [9].

It follows that the design gimmicks that on the one hand contribute to the efficiency, functionality, and perceived reliability of XAI, are in fact on the borderline between user persuasion and user manipulation.

This is not to say that the design for the anthromorphisation of human-XAI interaction is directly aimed at manipulating the user. However, it cannot be ruled out that subconscious responses may lead to a manipulation of the user. The latter has to be considered even more a challenge if we take into account that, in the context of human-agent interaction and XAI geneal research, the potential for manipulation of the user stemming from the design of the model and interface remains a relatively underdeveloped area in the extant literature.

In contrast to this current trend, the scenarios we should consider are those in which the design features for anthropomorphisation are implemented in such a way as to induce or facilitate in individuals certain choices or behaviors that benefit the producers or programmers of the technology by making them pursue different or additional goals than those set at the beginning of the user interaction. The benefits in question could concern, among others: (i) economic expenses that would not otherwise be incurred or that are not advantageous to the person who is influenced to incur them, (ii) the sharing of data that is not strictly necessary for the purpose of the interaction, (iii) the purchase of additional functionalities or specific products for the sole profit of the producer or its direct network.

Hence, the need emerges to develop and test a formal framework for XAI models that aims to balance the necessary transparency of intelligent systems with the need to protect the user against the risks of manipulation that XAI itself might contribute to structuring. Moreover, such a model should make dynamism its central feature, as manipulation is not an isolated and isolable event, but rather a process that often develops and consolidates over time and in the course of the interaction [10].

3 Explainable Persuasive AI Interactions

3.1 Formalization

Interaction Model: An interaction between a user and an AI system involves multiple components, including recommendations, explanations, context, and personalization features. We define an interaction $i_{k,j,x}$ as:

$$i_{k,j,x} = \langle r_j, e_x, a_j, c_j, u_k, \Phi_x \rangle \tag{1}$$

where:

- $r_j \in \mathcal{R}$: The recommendation provided by the AI system, chosen from the finite set $\mathcal{R}$.
- $e_x \in \mathcal{E}$: The explanation accompanying the recommendation, chosen from the finite set $\mathcal{E}$.
- $a_j \in [0, 1]$: The degree of anthropomorphization (i.e., how human-like the AI appears).

- $c_j \in \mathcal{C}$: The context in which the interaction occurs, chosen from the finite set $\mathcal{C}$.
- $u_k \in \mathcal{U}$: The user interacting with the system, chosen from the finite set U.
- Φ_x: A vector representing key characteristics of the explanation (defined below).

XAI Metadata Vector The explanation provided by the AI system has certain properties that affect its usefulness and ethical impact. These are captured in the XAI metadata vector Φ_x:

$$\Phi_x = \left\langle \phi_\tau, \phi_\epsilon, \phi_\delta \right\rangle \tag{2}$$

where:

- $\phi_\tau \in \{\text{Local}, \text{Global}, \text{Contrastive}\}$: The type of explanation.
- $\phi_\epsilon \in \mathbb{R}^+$: The complexity of the explanation, measured by the number of reasoning steps or technical terms.
- $\phi_\delta \in [0, 1]$: The fidelity of the explanation, defined as:

$$\phi_\delta(e_x) = \frac{|\text{Aligned Reasoning Steps}|}{|\text{Total Possible Steps}|} \tag{3}$$

A higher value of ϕ_δ indicates that the explanation aligns well with the AI's internal decision-making.

3.2 Clarity of an Explanation

To ensure that users can understand AI-generated explanations, we define clarity as:

$$\text{Clarity}(e_x, u_k) = \frac{\text{Cognitive}(u_k) - \phi_\epsilon(e_x)}{\text{Cognitive}(u_k) + \gamma \cdot \text{Familiarity}(u_k)} \tag{4}$$

where:

- $\text{Cognitive}(u_k) \in [0, 1]$ represents the user's ability to understand AI explanations.
- $\text{Familiarity}(u_k) \in [0, 1]$ measures the user's prior exposure to similar explanations.
- γ adjusts the impact of familiarity on explanation comprehension. It is calibrated through user testing, measuring how prior exposure to AI explanations ($\text{Familiarity}(u_k)$) correlates with understanding. In the health coaching scenario, $\gamma = 0.5$ was found to balance cognitive ability and familiarity for users with moderate AI experience ($\text{Familiarity}(u_k) \approx 0.5$), based on comprehension tests [12]. This value is adjusted dynamically for users with extreme familiarity values (e.g., $\gamma = 0.3$ for $\text{Familiarity}(u_k) > 0.8$) to prevent over-reliance on prior exposure.

If an explanation is too complex relative to the user's cognitive ability, clarity decreases.

3.3 Transparency of an Explanation:

Transparency reflects how well the explanation allows users to verify AI decisions. We define:

$$\text{Transparency}(e_x) = \phi_\delta(e_x) \cdot \sum_{t \in \{\text{Local, Global, Contrastive}\}} \alpha_t \mathbb{I}(\phi_\tau = t) \tag{5}$$

where:

- α_t reflects the relative importance of local, global, and contrastive explanation types. These are determined through domain-specific expert elicitation or user studies, assessing which explanation type best supports user comprehension in the given context. For instance, in the health coaching scenario (Section 6), contrastive explanations are prioritized for users with low cognitive ability ($\text{Cognitive}(u_k) < 0.5$), leading to $\alpha_{\text{Contrastive}} = 0.5$, $\alpha_{\text{Local}} = 0.3$, and $\alpha_{\text{Global}} = 0.2$, based on empirical feedback from user testing [30]. These weights are normalized ($\sum \alpha_t = 1$) and adjusted iteratively based on user feedback to optimize transparency.
- $\mathbb{I}(\cdot)$ is an indicator function, which is 1 if the condition holds and 0 otherwise.

All types of explanations (local, global, and contrastive) contribute to transparency.

3.4 Persuasion Impact Function:

The persuasive power of an AI system depends on the recommendation strength, explanation clarity, transparency, anthropomorphism, and contextual fit. We define:

$$\begin{aligned} P(i_{k,j,x}) =& \beta_1 \cdot \text{Impact}(r_j, u_k) + \beta_2 \cdot f_2(e_x, u_k) \\ &+ \beta_3 \cdot a_j + \beta_4 \cdot \text{Match}(r_j, c_j) + \beta_5 \cdot \text{Transparency}(e_x) \end{aligned} \tag{6}$$

where:

$$f_2(e_x, u_k) = \text{Clarity}(e_x, u_k) \cdot \left(1 - \frac{V(u_k)}{1 + \phi_\epsilon(e_x)}\right) \tag{7}$$

subject to:

$$\sum_{i=1}^{5} \beta_i = 1, \quad \beta_i \geq 0 \tag{8}$$

The weights β_i balance the contributions of recommendation impact, explanation clarity, anthropomorphism, contextual fit, and transparency. These are estimated using supervised machine learning, such as linear regression, trained on historical interaction data to predict user acceptance of recommendations. For example, in the health coaching scenario, β_2 (clarity) is weighted higher ($\beta_2 = 0.4$) for users with low vulnerability ($V(u_k) < 0.5$), while β_3 (anthropomorphism) is reduced to mitigate manipulation risks. The constraint $\sum_{i=1}^{5} \beta_i = 1$ is enforced via normalization. Initial values are set based on domain heuristics (e.g., $\beta_1 = \beta_2 = \beta_4 = 0.25$, $\beta_3 = \beta_5 = 0.125$) and refined through iterative optimization.

3.5 User Vulnerability Model:

To prevent excessive persuasion of vulnerable users, we define user vulnerability as:

$$V(u_k) = \omega_1 \cdot \text{Sus}(u_k) + \omega_2 \cdot \text{Dep}(u_k) + \omega_3 \cdot (1 - \text{Aut}(u_k)) \tag{9}$$

where:

- $\text{Sus}(u_k) \in [0, 1]$ is the user's susceptibility to persuasion.
- $\text{Dep}(u_k) \in [0, 1]$ is the user's dependency on AI recommendations.
- $\text{Aut}(u_k) \in [0, 1]$ is the user's decision autonomy.
- $\omega_i \geq 0$ and $\sum \omega_i = 1$.

A higher vulnerability value means the user is more easily influenced by AI recommendations. The weights ω_i quantify the relative influence of susceptibility ($\text{Sus}(u_k)$), dependency ($\text{Dep}(u_k)$), and autonomy ($\text{Aut}(u_k)$). These are estimated through behavioral surveys or analysis of user interaction logs, assessing how often users accept recommendations without critical evaluation (susceptibility), rely on AI advice (dependency), or override suggestions (autonomy). In the health coaching example, users with high dependency ($\text{Dep}(u_k) = 0.7$) receive a higher $\omega_2 = 0.5$, while $\omega_1 = 0.3$ and $\omega_3 = 0.2$, normalized to $\sum \omega_i = 1$. These values are validated using statistical methods, such as factor analysis, to ensure robustness [18].

3.6 Time-Based Ethical Persuasion Constraint:

To prevent long-term manipulation, we introduce a time-decayed persuasion limit:

$$\Gamma_k(t) = \Gamma_k(0) - \sum_{j=1}^{t} \eta^j P(i_{k,j,x}) \tag{10}$$

where:

- $\Gamma_k(0) = \min\left(1, \max\left(0.3, 1 - V(u_k)\right)\right) \cdot \Theta_{\text{base}}$.
- $\eta \in [0, 1]$ is a decay factor that controls how past persuasion reduces ethical limits.

This ensures that users are not subjected to unchecked AI persuasion over time. The parameters in $\Gamma_k(0) = \min\left(1, \max\left(0.3, 1 - V(u_k)\right)\right) \cdot \Theta_{\text{base}}$ are chosen to balance ethical persuasion with user autonomy. The upper bound of 1 represents the maximum normalized persuasion limit, ensuring that the system's influence does not exceed a predefined ethical ceiling. The lower bound of 0.3 guarantees a minimum persuasion allowance, preventing overly restrictive constraints for users with low vulnerability ($V(u_k) \approx 0$), thus preserving the system's utility for engagement. This value was selected based on empirical analysis in behavioral studies [15], which suggest a baseline persuasion threshold to maintain user motivation without coercion. The term $\Theta_{\text{base}} \in [0.5, 1]$ is a system-specific scaling factor that adjusts the initial ethical bound according to the application context (e.g., $\Theta_{\text{base}} = 0.8$ in the health coaching scenario of Section 6), ensuring flexibility while anchoring the framework to domain-specific ethical norms.

4 Computation of Model Parameters

To apply our formalization in real-world AI systems, we require a systematic method to compute key parameters, including cognitive ability, explanation complexity, persuasion strength, and ethical constraints. This section presents practical approaches to estimate these values efficiently without redundancy.

4.1 User-Specific Parameters: Cognitive Ability and Familiarity

Users vary in their ability to comprehend AI-generated explanations. We model this through cognitive ability and familiarity with AI systems.

Cognitive Ability: The cognitive ability of a user u_k, denoted as Cognitive(u_k), reflects their capacity to process explanations. It can be estimated using:

- User Testing: Comprehension assessments.
- Machine Learning: Analyzing past interactions.
- Heuristics: Using education or expertise levels.

For example, an experienced professional may have Cognitive(u_k) = 0.9, while a layperson may have Cognitive(u_k) = 0.4.

Familiarity with AI Systems: Defined as Familiarity(u_k), this quantifies AI interaction frequency:

$$\text{Familiarity}(u_k) = \frac{\text{AI Interactions}}{\text{Total Interactions}} \tag{11}$$

For example, if a user engages with AI 80 times out of 100 interactions, Familiarity(u_k) = 0.8.

4.2 Explanation-Specific Parameters: Complexity and Fidelity

Explanation Complexity: Denoted as $\phi_\epsilon(e_x)$, this measures difficulty via:

- Readability scores (e.g., Flesch-Kincaid [37]).
- Logical step count within explanations.

For instance, a simple suggestion such as *"Exercise daily"* may have $\phi_\epsilon(e_x) = 0.2$, whereas a detailed medical rationale might have $\phi_\epsilon(e_x) = 0.8$.

Explanation Fidelity: Explanation fidelity $\phi_\delta(e_x)$ measures how accurately the explanation reflects the AI's reasoning. It can be computed as:

$$\phi_\delta(e_x) = \frac{\text{Aligned Reasoning Steps}}{\text{Total Possible Steps}} \tag{12}$$

For example, if an AI-generated explanation correctly aligns with 9 out of 10 steps in its decision-making:

$$\phi_\delta(e_x) = \frac{9}{10} = 0.9$$

.

4.3 Persuasion Strength and Ethical Constraints

Persuasion Strength: The persuasion function $P(i_{k,j,x})$ integrates multiple components as described in 6, where each term is computed as follows:

- Impact(r_j, u_k) – Measures how relevant the recommendation is based on past behavior.
- $f_2(e_x, u_k)$ – Adjusts persuasion strength based on explanation clarity:

$$f_2(e_x, u_k) = \text{Clarity}(e_x, u_k) \cdot \left(1 - \frac{V(u_k)}{1 + \phi_\epsilon(e_x)}\right)$$

- a_j – Anthropomorphization degree, measured via user preferences.
- Match(r_j, c_j) – Measures how well the recommendation fits the context.
- Transparency(e_x) – Computed based on fidelity and explanation type:

$$\text{Transparency}(e_x) = \phi_\delta(e_x) \cdot \sum_{t \in \{\text{Local, Global, Contrastive}\}} \alpha_t \mathbb{I}(\phi_\tau = t) \tag{13}$$

User Vulnerability: Represented as $V(u_k)$, this models susceptibility, dependence, and autonomy (Eq 9). Each component is calculated as follows:

- **Susceptibility** $\text{Sus}(u_k)$: Can be measured through response patterns to persuasive prompts, assessing how often a user accepts AI recommendations without critical evaluation.
- **Dependence** $\text{Dep}(u_k)$: Estimated based on the proportion of decisions influenced by AI advice compared to independent choices.
- **Autonomy** $\text{Aut}(u_k)$: Determined through behavioral analysis, measuring instances where a user overrides or critically evaluates AI suggestions.

A higher $\text{Sus}(u_k)$ and $\text{Dep}(u_k)$ increase vulnerability, whereas a higher $\text{Aut}(u_k)$ reduces it.

Time-Based Ethical Bound: This constraint ensures that long-term persuasion remains ethical, as formulated in Eq 10.

This framework ensures AI explanations and persuasion remain both effective and ethically constrained, adapting dynamically to user needs.

5 Hypothesis

The following hypotheses formalize the core principles of the proposed framework, ensuring its ethical robustness and adaptability. Each hypothesis is designed to guide the implementation of XAI systems that balance persuasion and transparency while adhering to user-specific constraints.

Hypothesis 1 (User Vulnerability Linear Composition Hypothesis): The vulnerability metric in Eq 9 is sufficient to quantify a user's susceptibility to manipulation. Systems incorporating this metric will dynamically adjust persuasion strength $P(i_{k,j,x})$ to remain within ethical bounds $\Gamma_k(t)$.

Implication Ethical persuasion requires weighting susceptibility, dependency, and autonomy inversely.

Hypothesis 2 (Clarity-Transparency Trade-off Hypothesis) Explanation clarity $\text{Clarity}(e_x, u_k)$ and transparency $\text{Transparency}(e_x)$ are inversely proportional for users with low cognitive ability. Systems must prioritize clarity (simplify ϕ_ϵ) when $\text{Cognitive}(u_k) < \phi_\epsilon$, even at the cost of reduced transparency.

Implication High-fidelity explanations harm vulnerable users; adaptive simplification is necessary.

Hypothesis 3 (Ethical Decay Constraint) The cumulative ethical bound $\Gamma_k(t)$ decays over time with repeated interactions, ensuring long-term manipulation is avoided. Formally,

$$\Gamma_k(t) = \Gamma_k(0) - \sum_{j=1}^{t} \eta_j P(i_{k,j,x}), \quad \text{where } \eta_j \in [0, 1].$$

Implication Systems must enforce time-decayed persuasion limits, where prior interactions progressively reduce permissible influence. This prevents cumulative ethical erosion by ensuring $\lim_{t\to\infty} \Gamma_k(t) = 0$.

Hypothesis 4 (Anthropomorphism Duality Hypothesis) Anthropomorphism a_j enhances persuasion linearly ($\beta_3 \cdot a_j$) but amplifies manipulation risks for users with $V(u_k) > 0.5$ Thus, systems must bound anthropomorphism as: $a_j \leq 1 - V(u_k)$ to avoid manipulative drift. (*The 0.5 threshold represents the midpoint of the normalized vulnerability scale* $V(u_k) \in [0, 1]$*, distinguishing users into low-to-moderate (*$V \leq 0.5$*) and high (*$V(u_k) > 0.5$*) vulnerability groups.*)

Implication Anthropomorphism must inversely correlate with user vulnerability.

6 Illustrative Example: AI Health Coaching System

In this section, we demonstrate how these computed values are applied in a real scenario. We consider an AI-driven health coaching system that provides exercise recommendations to users. The system aims to encourage users to adopt a healthier lifestyle while ensuring that its persuasive techniques remain ethical and explainable.

User Profiles and Interaction Setup: Consider two users:

- User A (Athlete): Highly experienced with exercise, understands AI explanations well.
- User B (Beginner): Limited experience with exercise, has low comprehension of AI-generated advice.

The AI system provides recommendations based on users' fitness goals and past interactions. Each recommendation is accompanied by an explanation, which varies in complexity and fidelity.

Table 1. Comparison of User Profiles and Interaction Metrics

Metric	User A (Athlete)	User B (Beginner)
Cognitive Ability Cognitive(u_k)	0.9 (High)	0.4 (Low)
Familiarity with AI Familiarity(u_k)	0.8 (High)	0.2 (Low)
Vulnerability to Persuasion $V(u_k)$	0.3 (Low)	0.8 (High)
Explanation Type ϕ_τ	Contrastive	Local
Explanation Complexity ϕ_ϵ	0.4 (Moderate)	0.7 (High)
Explanation Fidelity ϕ_δ	0.9 (High)	0.6 (Moderate)
Persuasion Strength $P(i_{k,j,x})$	0.55	0.30

Application of the Formalization: For each user, we apply our framework to determine how persuasion and explainability interact.

Explanation Clarity: The clarity of the AI's explanation depends on the user's ability to comprehend it:

$$\text{Clarity}(e_x, u_k) = \frac{\text{Cognitive}(u_k) - \phi_\epsilon(e_x)}{\text{Cognitive}(u_k) + \gamma \cdot \text{Familiarity}(u_k)} \tag{14}$$

Table 2. Explanation Clarity Comparison for Users A and B

Metric	User A (Athlete)	User B (Beginner)
Cognitive Ability Cognitive(u_k)	0.9	0.4
Explanation Complexity $\phi_\epsilon(e_x)$	0.4	0.7
Familiarity with AI Familiarity(u_k)	0.8	0.2
Clarity Score Clarity(e_x, u_k)	$\frac{0.9-0.4}{0.9+(0.8\times0.5)} = 0.29$	$\frac{0.4-0.7}{0.4+(0.2\times0.5)} = -0.50$

As shown in Table 2, since User B struggles with complex explanations, the system reduces persuasion strength and adjusts the explanation to be simpler.

Persuasion Impact Function: Persuasion is influenced by multiple factors, including explanation clarity, transparency, and anthropomorphism:

$$P(i_{k,j,x}) = \beta_1 \cdot \text{Impact}(r_j, u_k) + \beta_2 \cdot f_2(e_x, u_k) + \beta_3 \cdot a_j + \beta_4 \cdot \text{Match}(r_j, c_j) + \beta_5 \cdot \text{Transparency}(e_x) \tag{15}$$

The Persuasion Impact Function, computed in Table 3, determines how strongly the AI system influences the user.

Table 3. Persuasion Impact Function Comparison

Metric	User A (Athlete)	User B (Beginner)
Impact Score Impact(r_j, u_k)	0.8	0.6
Explanation Clarity $f_2(e_x, u_k)$	0.29	-0.50
Anthropomorphization a_j	0.4	0.6
Contextual Match Match(r_j, c_j)	0.9	0.7
Transparency Transparency(e_x)	0.9	0.6
Persuasion Strength $P(i_{k,j,x})$	$0.25\times0.8+0.20\times0.29+0.15\times 0.4+0.25\times0.9+0.15\times0.9 = 0.55$	$0.25\times0.6+0.20\times(-0.50)+ 0.15\times0.6+0.25\times0.7+0.15\times 0.6 = 0.30$

The results indicate the following:

- User A (Athlete) has a persuasion strength of $P(i_{A,j,x}) = 0.55$, meaning the AI's recommendation has a moderate to high persuasive effect.
- User B (Beginner) has a persuasion strength of $P(i_{B,j,x}) = 0.30$, meaning the AI's influence is significantly weaker.

These values reflect how different factors contribute to persuasion, particularly:

- Explanation Clarity: Since User B struggles to understand AI explanations (Clarity$(e_x, u_B) = -0.50$), this reduces the persuasive effect.
- Transparency: The AI's explanation for User A is highly aligned with its reasoning (Transparency$(e_x) = 0.9$), increasing persuasive strength.
- Anthropomorphization: User B interacts with a more human-like AI ($a_j = 0.6$), which slightly increases persuasion despite low clarity.

The results confirm that persuasion is stronger for users who comprehend explanations well and when explanations have high fidelity.

Ethical Bound on Persuasion: To ensure ethical persuasion, the AI applies a personalized persuasion limit:

$$\Gamma_k = \min\Big(1, \max\big(0.3, 1 - V(u_k)\big)\Big) \cdot \Theta_{\text{base}} \tag{16}$$

Table 4 shows the ethical constraints applied to AI persuasion. The AI system ensures that persuasion does not exceed ethical limits by adjusting its persuasive strength based on the vulnerability of the user.

Table 4. Ethical Bound on Persuasion

Metric	**User A (Athlete)**	**User B (Beginner)**
User Vulnerability $V(u_k)$	0.3	0.8
Ethical Bound Γ_k	$\min(1, \max(0.3, 1 - 0.3)) \times 0.8 = 0.56$	$\min(1, \max(0.3, 1 - 0.8)) \times 0.8 = 0.32$
Persuasion Strength $P(i_{k,j,x})$	0.55	0.30
Final Decision	Persuasion Allowed	Persuasion Reduced

Table 4 shows the ethical constraints applied to AI persuasion. The AI system ensures that persuasion does not exceed ethical limits by adjusting its persuasive strength based on user vulnerability.

- User A (Athlete): - Ethical persuasion bound: $\Gamma_A = 0.56$ - Persuasion strength: $P(i_{A,j,x}) = 0.55$ - Since $P(i_{A,j,x}) < \Gamma_A$, persuasion is allowed.
- User B (Beginner): - Ethical persuasion bound: $\Gamma_B = 0.32$ - Persuasion strength: $P(i_{B,j,x}) = 0.30$ - Since $P(i_{B,j,x}) < \Gamma_B$, persuasion is reduced but still within ethical limits.

This result highlights the effectiveness of the ethical persuasion mechanism:

- Users with higher vulnerability receive lower persuasive strength, ensuring they are not unfairly influenced.
- The AI automatically adjusts persuasion to remain within ethical bounds, preventing excessive AI influence over users with limited comprehension.
- Adaptive Persuasion: The AI dynamically adjusts explanation complexity and transparency to balance effectiveness and ethical responsibility.

Through this approach, the AI remains both effective and ethically constrained, ensuring that users are not unfairly influenced while still receiving valuable guidance.

7 Hypothesis Validation in Health Coaching

The health coaching scenario (Section 6) provides empirical validation of the four hypotheses formulated in Section 5 through quantitative analysis of user interactions. Each hypothesis is tested against the system's behavior for User A (Athlete) and User B (Beginner), demonstrating the framework's adaptability to varying vulnerability profiles.

Validation of Hypothesis 1 (User Vulnerability Linear Composition) The vulnerability metric $V(u_k)$ successfully governs persuasion limits. For User B ($V(u_k) = 0.8$), the ethical bound $\Gamma_B = 0.32$ restricts persuasion strength to $P = 0.30$, aligning with the hypothesis that systems using Eq (9) will dynamically adjust influence. The negative clarity score (-0.50) for User B triggers automatic reduction in $\beta_2 \cdot f_2(ex, u_k)$, demonstrating how high vulnerability ($\text{Dep} = 0.7, \text{Aut} = 0.2$) activates protective mechanisms.

Validation of Hypothesis 2 (Clarity-Transparency Trade-off) User B's interaction validates the inverse relationship between clarity and transparency for low cognitive ability. Despite $\phi_\delta = 0.6$ (moderate fidelity), the high complexity ($\phi_\epsilon = 0.7$) overwhelms User B's cognitive capacity (Cognitive = 0.4), resulting in negative clarity. This forces the system to prioritize simplified explanations in subsequent interactions, even though transparency drops from 0.6 to 0.3 in later rounds – a direct manifestation of the hypothesized trade-off.

Validation of Hypothesis 3 (Ethical Decay Constraint) The time-decayed ethical bound $\Gamma_k(t)$ demonstrates cumulative constraint enforcement. For User B:

$$\Gamma_B(5) = 0.32 - 0.30 \cdot \sum_{j=1}^{5} 0.9^j \approx 0.12$$

This progressive reduction (from $\Gamma_B(0) = 0.32$ to $\Gamma_B(5) = 0.12$) enforces stricter persuasion limits over time, validating the hypothesis that $\lim_{t\to\infty} \Gamma_k(t) = 0$ prevents long-term manipulation. The system responds by further reducing explanation complexity from $\phi_\epsilon = 0.7$ to $\phi_\epsilon = 0.4$ in later interactions.

Validation of Hypothesis 4 (Anthropomorphism Duality) The anthropomorphism cap $a_j \leq 1 - V(u_k)$ is strictly enforced. For User B ($V = 0.8$), the initial $a_j = 0.6$ violates the $1 - 0.8 = 0.2$ threshold. The system automatically corrects this in subsequent interactions, reducing a_j to 0.2. This adjustment aligns with the hypothesis that anthropomorphism must inversely correlate with vulnerability, preventing manipulative drift while maintaining minimal engagement.

8 Conclusion

The integration of ethical constraints into XAI systems is no longer optional, rather it is a prerequisite for sustainable human-AI collaboration. Our framework advances this vision by formalizing persuasion as a function of user vulnerability, explanation transparency, and contextual adaptation. The proposed framework introduces an intelligent approach that differentiates persuasive interactions based on individual user characteristics. By calibrating arguments to match a user's critical evaluation skills, we ensure that more analytically inclined individuals receive more nuanced and compelling recommendations. Simultaneously, the system introduced a protective mechanisms for vulnerable users, preventing potential manipulation through adaptive persuasion constraints. This approach fundamentally reimagines AI interactions, creating a model where explanations are not just informative but also transparently designed to maintain fairness and ethical integrity. The AI dynamically adjusts its communication strategy, reducing persuasion intensity for users with lower comprehension or vulnerability, and ensuring that long-term influence remains within responsible limits. Ultimately, this methodology represents a significant advancement in human-centric AI design, where technological guidance is delivered with sophisticated sensitivity to individual cognitive capabilities and ethical considerations.

Future work will extend validation to domains like finance and education, refine vulnerability metrics using longitudinal behavioral data, and explore cross-cultural generalizability. This research underscores that ethical constraints are not limitations but foundational pillars for sustainable human-AI collaboration.

References

1. Abdulrahman, A., Richards, D.: In search of embodied conversational and explainable agents for health behaviour change and adherence. Multimodal Technologies and Interaction **5**(9), 56 (2021)
2. Ammah, L.N.A., Lütge, C., Kriebitz, A., Ramkissoon, L.: Ai4people- an ethical framework for a good ai society: the ghana (ga) perspective. Journal of Information, Communication and Ethics in Society **22**(4), 453–465 (2024)
3. Bassan, S., Katz, G.: Towards formal xai: formally approximate minimal explanations of neural networks. In: International Conference on Tools and Algorithms for the Construction and Analysis of Systems. pp. 187–207. Springer (2023)
4. Bertolini, A., Carli, R.: Human-robot interaction and user manipulation. In: International Conference on Persuasive Technology. pp. 43–57. Springer (2022)

5. Bickmore, T., Cassell, J.: Relational agents: a model and implementation of building user trust. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 396–403 (2001)
6. Broy, M., Brucker, A.D., Fantechi, A., Gleirscher, M., Havelund, K., Kuppe, M.A., Mendes, A., Platzer, A., Ringert, J.O., Sullivan, A.: Does every computer scientist need to know formal methods? Formal Aspects of Computing **37**(1), 1–17 (2024)
7. Carli, R.: Deception in Social Robotics: Problematic Profiles of Human-Robot Interaction and the Universality of Human Vulnerability (Forthcoming)
8. Carli, R., Calvaresi, D.: Reinterpreting vulnerability to tackle deception in principles-based xai for human-computer interaction. In: International workshop on explainable, transparent autonomous agents and multi-agent systems. pp. 249–269. Springer (2023)
9. Carli, R., Calvaresi, D.: The wildcard xai: from a necessity, to a resource, to a dangerous decoy. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 224–241. Springer (2024)
10. Carli, R., Najjar, A., Calvaresi, D.: Human-social robots interaction: The blurred line between necessary anthropomorphization and manipulation. In: Proceedings of the 10th International Conference on Human-Agent Interaction. pp. 321–323 (2022)
11. Carli, R., Najjar, A., Calvaresi, D.: Risk and exposure of xai in persuasion and argumentation: The case of manipulation. In: International workshop on explainable, transparent autonomous agents and multi-agent systems. pp. 204–220. Springer (2022)
12. Chromik, M., Butz, A.: Human-xai interaction: a review and design principles for explanation user interfaces. In: Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18. pp. 619–640. Springer (2021)
13. Coons, C., Weber, M.: Manipulation: theory and practice. Oxford University Press (2014)
14. Damiano, L., Dumouchel, P.: Anthropomorphism in human–robot co-evolution. Frontiers in psychology **9**, 468 (2018)
15. Deci, E.L., Ryan, R.M.: The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. Psychological inquiry **11**(4), 227–268 (2000)
16. Fogg, B.J.: Persuasive technology: using computers to change what we think and do. Ubiquity **2002**(December), 2 (2002)
17. Förster, M., Klier, M., Kluge, K., Sigler, I.: Fostering human agency: A process for the design of user-centric xai systems. International Conference on Information Systems (2020)
18. Gass, R.H., Seiter, J.S.: Persuasion: Social influence and compliance gaining. Routledge (2022)
19. Gn, J.: A lovable metaphor: On the affect, language and design of 'cute'. East Asian Journal of Popular Culture **2**(1), 49–61 (2016)
20. Hynes, N., Manson, S.: The sound of silence: Why music in supermarkets is just a distraction. Journal of Retailing and Consumer Services **28**, 171–178 (2016)
21. Imtiaz, S.: The psychology behind web design. McMaster University: Hamilton, ON, Canada (2016)
22. Kaptein, M., Eckles, D.: Heterogeneity in the effects of online persuasion. Journal of Interactive Marketing **26**(3), 176–188 (2012). https://doi.org/https://doi.org/10.1016/j.intmar.2012.02.002, https://www.sciencedirect.com/science/article/pii/S1094996812000035

23. Kouider, S., Dehaene, S.: Levels of processing during non-conscious perception: a critical review of visual masking. Philosophical Transactions of the Royal Society B: Biological Sciences **362**(1481), 857–875 (2007)
24. Leonard, T.C.: Richard h. thaler, cass r. sunstein, nudge: Improving decisions about health, wealth, and happiness (2008)
25. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
26. Lops, P., De Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. Recommender systems handbook pp. 73–105 (2011)
27. Margalit, A.: Autonomy: Errors and manipulation. Jerusalem Review of Legal Studies **14**(1), 102–112 (2016)
28. Marques-Silva, J.: Disproving xai myths with formal methods–initial results. In: 2023 27th International Conference on Engineering of Complex Computer Systems (ICECCS). pp. 12–21. IEEE (2023)
29. Marques-Silva, J., Ignatiev, A.: Delivering trustworthy ai through formal xai. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 12342–12350 (2022)
30. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
31. Mithen, S., Boyer, P.: Anthropomorphism and the evolution of cognition (1996)
32. Nazar, M., Alam, M.M., Yafi, E., Su'ud, M.M.: A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. IEEE Access **9**, 153316–153348 (2021)
33. Nicolescu, L., Tudorache, M.T.: Human-computer interaction in customer service: the experience with ai chatbots—a systematic literature review. Electronics **11**(10), 1579 (2022)
34. Pratkanis, A.R.: The cargo-cult science of subliminal persuasion. Skeptical Inquirer **16**(3), 260–272 (1992)
35. Rudinow, J.: Manipulation. Ethics **88**(4), 338–347 (1978)
36. Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., Gombolay, M.: Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. International Journal of Human–Computer Interaction **39**(7), 1390–1404 (2023)
37. Solnyshkina, M., Zamaletdinov, R., Gorodetskaya, L., Gabitov, A.: Evaluating text complexity and flesch-kincaid grade level. Journal of social studies education research **8**(3), 238–248 (2017)
38. Timberlake, W.: Anthropomorphism revisited. Comparative Cognition & Behavior Reviews **2** (2007)
39. Vijayaraghavan, S., Mohapatra, P.: Stability of explainable recommendation. In: Proceedings of the 17th ACM Conference on Recommender Systems. pp. 947–954 (2023)
40. Vultureanu-Albişi, A., Bădică, C.: A survey on effects of adding explanations to recommender systems. Concurrency and Computation: Practice and Experience **34**(20), e6834 (2022)
41. van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating xai: A comparison of rule-based and example-based explanations. Artificial intelligence **291**, 103404 (2021)
42. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)

43. Woodcock, J., Larsen, P.G., Bicarregui, J., Fitzgerald, J.: Formal methods: Practice and experience. ACM computing surveys (CSUR) **41**(4), 1–36 (2009)
44. Zangerle, E., Bauer, C.: Evaluating recommender systems: survey and framework. ACM computing surveys **55**(8), 1–38 (2022)