Preference-Based

Inverse Reinforcement Learning

for ***Emotional Companionship Robots***

——Isabella's 60-Day Emotional Trajectory

Wanchen Lian 50018021 AI Thrust

Problem Statement: When Mood Scores *Deceive*

The Paradox

Traditional companion robots pursue a simple goal: **maximize** Isabella's daily happiness (mood scores 1-10).

They observe:

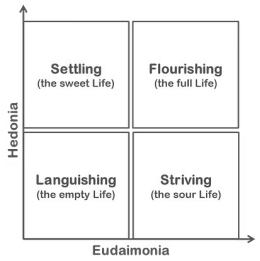
- **Sunday scrolling**: 8/10 mood ✓ → Robot reinforces this pattern
- **Exhibition work week**: 6/10 mood ☒ → Robot suggests avoidance
- **Two months later**, Isabella reflects: *"I feel empty. I don't want more 8/10 days; I want more 6/10 days like that week."*

Why did maximizing mood fail?

Because mood scores combine **Hollow Pleasure** (Hedonia) with **Rich Meaning** (Eudaimonia).

Layer	Captures	Misses	Result
Immediate State	What Isabella feels now(Mood)	What she actually values	❌ Noise
Stable Pattern	What Isabella fundamentally wants(Personality)	Reflected in long-term behavior	✅ Signal

The Real Question is not *"How do we maximize her mood?"*,
but ***"What is Isabella's emotional preference pattern?"***
Why she genuinely prefers "6/10 meaningful" over "8/10 hollow"



- Mood score: Moment-to-moment emotional state (hedonic)
- Personality: Stable value structure (eudaimonic + individual trait)[*Ryan & Deci, 2001; Kahneman, 2005*]

- Mood scores are noise. Personality is signal.

Why IRL: From Observation to *Preference*

The Core Problem

We observe Isabella's choices (what she does), but need to infer her rewards (why she values it).

Cannot use: Simple statistics, clustering, surveys — they cannot reverse-engineer hidden preferences.

What We Get

$$R_\theta(s) = \theta^\top \phi(s)$$

A weekly emotional preference function:

- where ϕ(s) is an interpretable feature vector for a week (valence, stability, social / work / rest, conflict, story phase).

For the Robot

- Interpretation of θ (example axes):
  - θ_valence_mean ≈ how much she values "feeling good this week"
  - θ_social_support ≈ how much she values warm, supportive social contact
  - θ_structured_work ≈ how much she values meaningful, structured busy weeks
  - θ_conflict / θ_election ≈ how much she dislikes emotionally draining, conflict-heavy weeks
  - θ_stability ≈ how much she prefers emotional stability over volatility

Not: Maximize mood; **But:** Optimize for *her* θ

Why IRL Works: Three Reasons

Hidden Rewards in Choices

- Rewards are implicit in long-term choices, not explicitly given.
- IRL reverses the direction: Behavior / emotional patterns → inferred Reward R_θ(s)

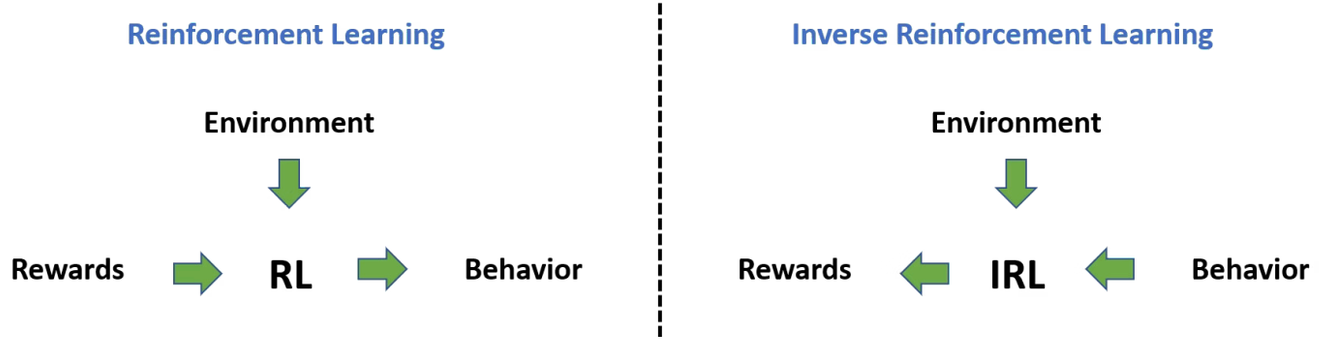
Temporal Structure

- Emotions are not independent moments; they form a 60-day trajectory.
- After learning R_θ(s), we treat the 54 weekly states as a Markov Reward Process and define a value function:
  - $V_\theta(t) = \Sigma_{k \geqslant t} \gamma^{k-t} \cdot R_\theta(s_k)$
- Value functions naturally capture the cumulative value of this emotional week
- IRL discount rates encode: "why she prefers struggle-now over comfort-now"

Personality Alignment

- The learned θ parameters tell us which dimensions she values or avoids:
  - e.g., social vs conflict, structured work vs exhausting pressure, stability vs chaos.
- We can validate:

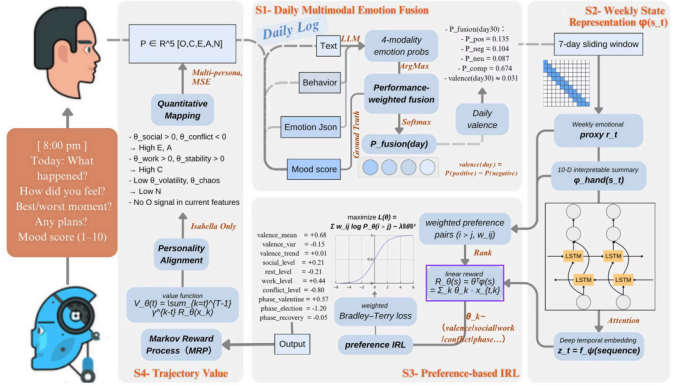
"Does this preference pattern match her Big Five profile and persona narrative?"



Methodology

Background

- Environment: 60-day run in a Stanford “Generative Agents” small-town–style simulation (Smallville-inspired).
- Agent: Isabella, one town agent with a fixed persona and Big-5 profile, placed into a Valentine / art-show / election / recovery storyline.
- Data: for each of 60 days we log behaviors.



1. **Daily Fusion** → Aggregate 4 modalities (mood, text, behavior, context) into daily emotion distribution
2. **Weekly Aggregation** → Compress 7-day windows into feature vectors ϕ(s_t)
3. **IRL Learning** → Extract implicit preference pairs, learn weights θ via Bradley-Terry
4. **Value Embedding** → Map θ into MRP to compute emotional trajectory value V(t)
5. **Validation** → Align θ with Big Five personality to verify consistency

Input & Output

Input: 60 days of multimodal data (mood scores, behavior logs, reflections, dialogue)

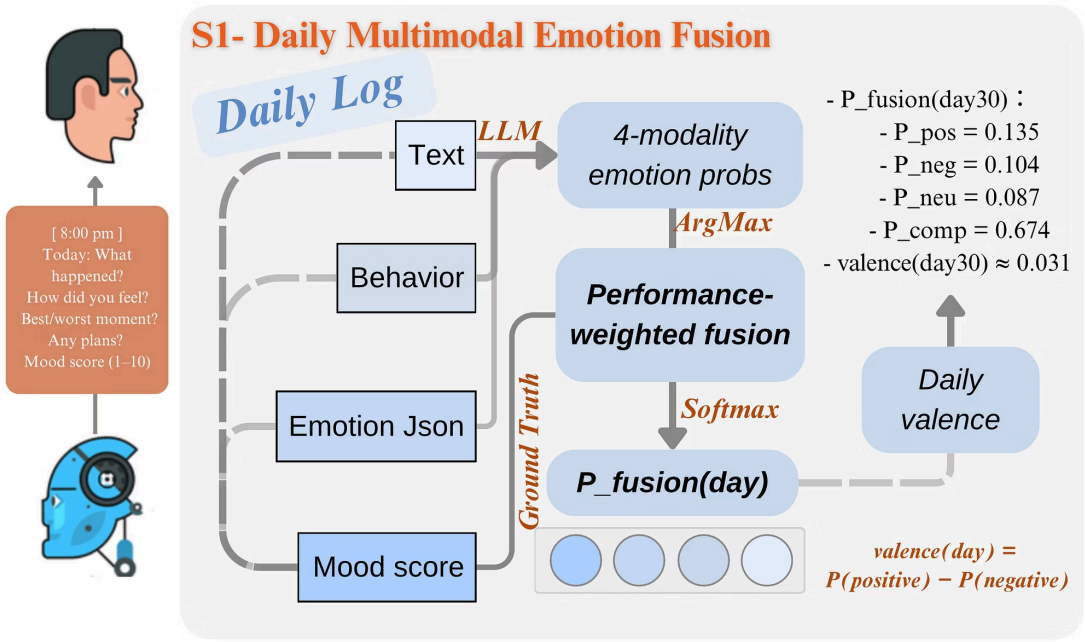
Output:

- θ (weekly preference weights revealing what she truly values)
- V(t) (emotional value function showing narrative arc)
- Big Five alignment (personality consistency check)

Why This Pipeline?

- S1: Convert messy daily signals into clean weekly representations
- S2: Infer hidden preferences from observed behavior
- S3: Capture temporal dependencies (not just static preferences)
- S4: Ensure learning is grounded in stable personality traits

Stage 1: Daily Multimodal Fusion



Isabella leaves behind four daily signals, each telling a partial truth:

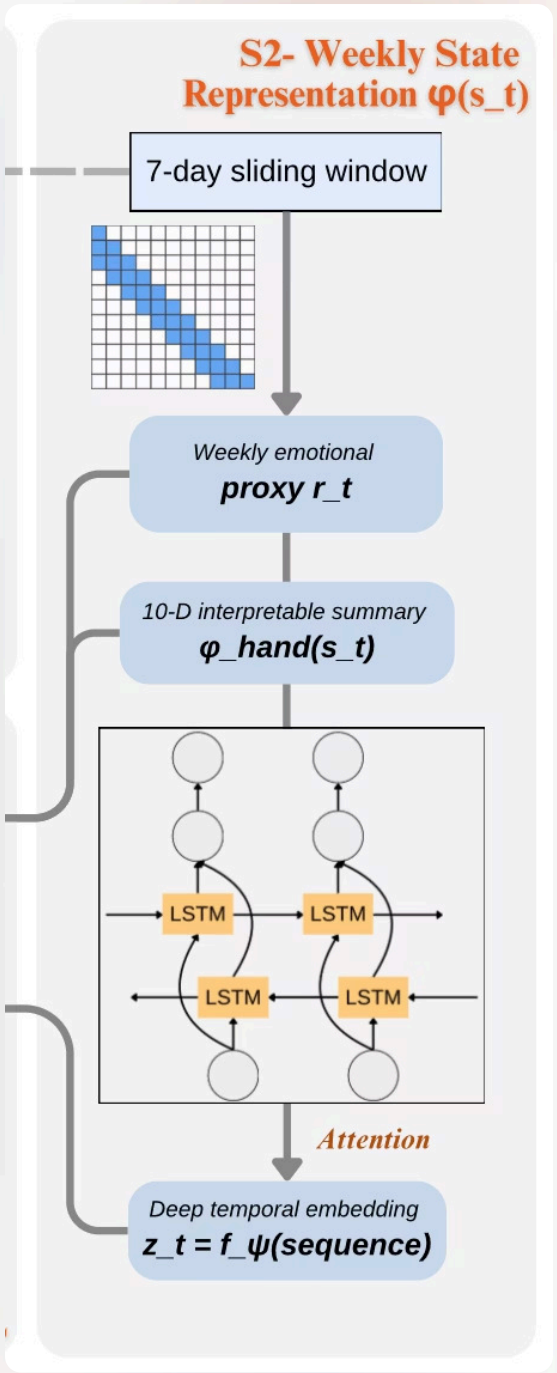
- What she writes in her evening diary
- How she spent her time (social, work, rest, creative)
- What emotion label she assigns and why
- Her numerical mood score (1–10)

The Process

1. **Extract four modality predictions**: Each produces a probability distribution over four emotion classes
2. **Compute per-modality accuracy**: Validate on held-out data
3. **Softmax-weight the accuracies**: $w_i = \exp(\text{accuracy}_i) / \Sigma \exp(\text{accuracy}_j)$
4. **Fuse**: $P_{\text{fusion}}(\text{day}) = \Sigma\, w_i \times P_i(\text{day})$
5. **Extract valence**: $\text{valence}(\text{day}) = P(\text{positive}) - P(\text{negative})$

The Output

A single, calibrated emotional signal for each of 60 days: a 4-dimensional probability distribution $P_{\text{fusion}}(\text{day})$ representing Isabella's true emotional state that day, plus a scalar valence curve (positive or negative) that becomes the foundation for all downstream learning.



Stage 2: Weekly State Representation & Deep Temporal Embedding

The 7-Day Sliding Window

We create 54 overlapping weeks by sliding a 7-day window across the 60 days (week 0 = days 0–6, week 1 = days 1–7, and so on).

Building the Interpretable Feature Vector ϕ(s_t)

For each week t, we compute a ~10-dimensional feature vector that captures three key dimensions:

1

Emotional Dynamics

- Mean valence across the 7 days
- Variance (emotional stability or turbulence?)
- Trend (improving, declining, or flat?)

2

Behavioral Balance

- Average social time
- Work intensity
- Rest and recovery
- Conflict frequency

3

Story Context

- One-hot encoding for life phase (e.g., Valentine prep, art exhibition, election chaos, recovery period)

The Deep Temporal Embedding z_t

In parallel, we feed the rolling 7-day valence sequence into a BiLSTM with an attention mechanism. This neural network learns hidden temporal patterns that raw features might miss—for example, the significance of a low-valence day following a high-valence week, or cyclical emotional rhythms.

Combining Both Representations

The final weekly state x_t can be used as just ϕ(s_t) for interpretability, or we can concatenate z_t for richer learning: x_t = [ϕ(s_t) ; z_t].

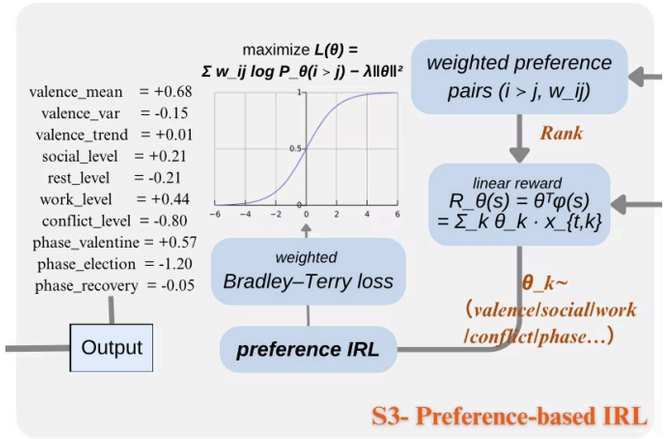
This gives us both transparency (what features matter) and expressiveness (what temporal patterns matter).

The Output

- 54 weekly states with interpretable features
- 54 deep temporal embeddings
- A unified representation ready for the IRL layer

Stage 3: Preference-Based IRL

We never tell the model: "this week = 0.73, that week = 0.21." Instead, we only say: "this week clearly felt better than that week." From these comparisons, the model learns: what kind of weeks Isabella really prefers



Learning θ: Uncovering the Weight Vector

Each week t has a feature vector φ(s_t) = [pleasure_mean, engagement_level, stress, recovery, challenge, ...] We want to find weights θ such that **preferred weeks score higher than non-preferred weeks:**

$$P(\text{week } i \succ \text{ week } j) = \sigma(\theta^T \phi(s_i) - \theta^T \phi(s_j))$$

This is the **Bradley-Terry preference model**—a standard way to rank items from pairwise preferences.

We maximize log-likelihood over all preference pairs:

$$\max_\theta \sum_{i \succ j} \log P(i \succ j) - \lambda \|\theta\|^2$$

The L2 regularization implements **Maximum Entropy**: "Among all θ that fit the observed preferences, choose the simplest one—don't overfit."

Is This Just Logistic Regression?

- Logistic regression: predict labels.
- Preference-based IRL: use a logistic model on preferences to recover a reward function for RL.

Input Data

- **Weekly state vectors (x_t)**: Emotional level, stability, social/work/rest balance, conflict frequency, story phase.
- **Weekly proxy scores (r_t)**: Mean valence over 7 days, solely used to **rank weeks**.

Step 1: Infer Pairwise Preferences

For every pair of weeks (i, j): Compare: r_i vs r_j

- If **r_i** is clearly higher than **r_j** (above a margin), we create a preference: week i ☒ week j
- Preference weight ∝ (r_i − r_j)

Example:

- Valentine (r = 0.77) ☒ Election (r = 0.13) → strong preference
- Week 5 (r = 0.65) ☒ Week 8 (r = 0.58) → weak preference

Output from Step 1: ~30 preference pairs.

Step 2: Learn Reward Function (θ)

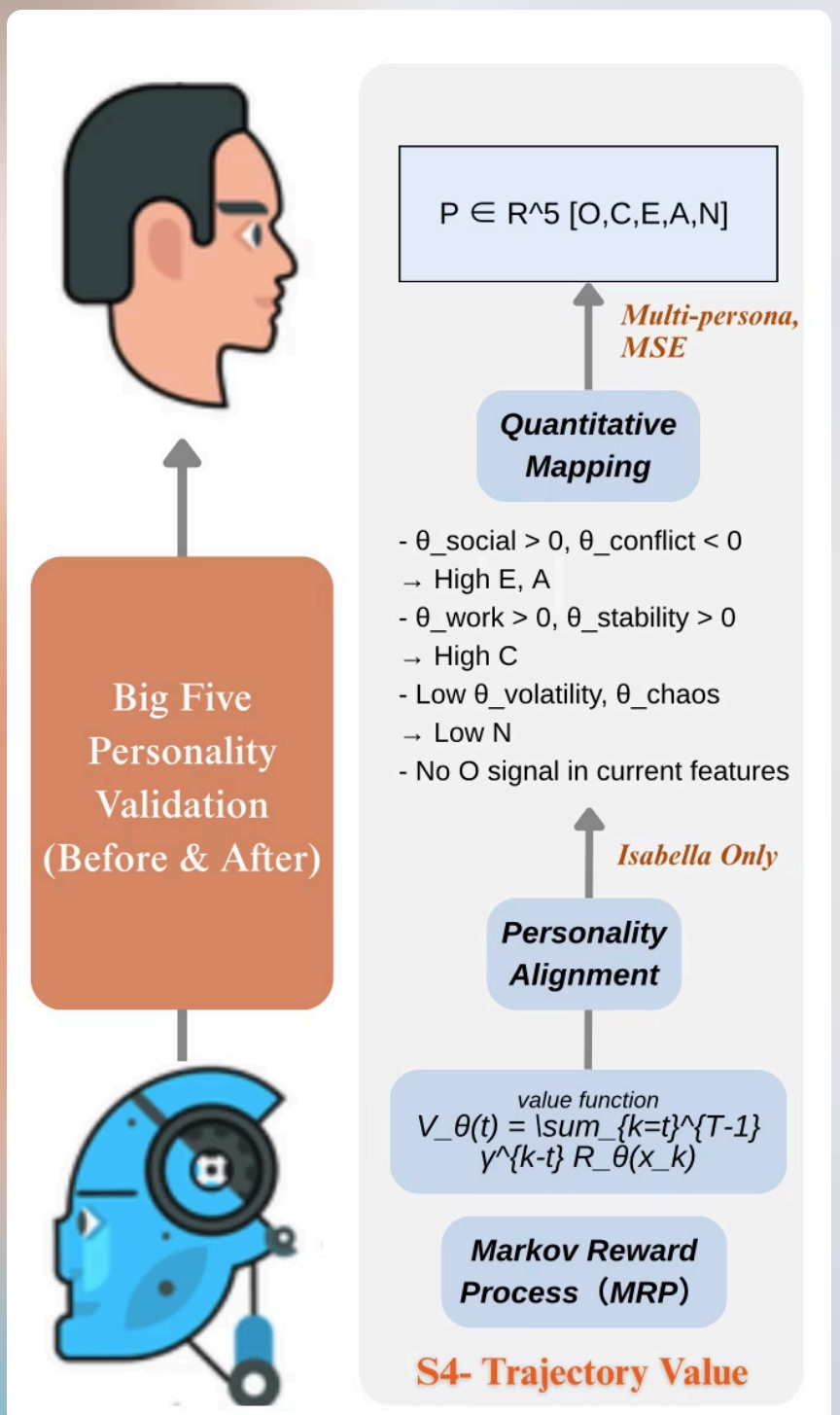
We assume a linear reward function: R_θ(x_t) = θᵀφ(x_t), where θ represents feature weights.

We train θ using a Bradley-Terry / logistic preference loss.

Optimization: Maximize log-likelihood over all preference pairs & avoid overfitting

Output: Discovered Preferences

- **Weight vector (θ)**: Reveals which weekly patterns Isabella rewards or penalises (e.g., negative weight on conflict, positive on structured work).
- **IRL reward score (R_θ(x_t))**: A value for each of the 54 weeks, reflecting its worth in Isabella's true internal preference space.



Stage 4: Trajectory Value & Personality Alignment

S3 gives us "how good each week itself is" ($R_\theta(x_t)$). In S4, we ask a longer-horizon question: "If life starts from this week, how good does the whole future stretch look?" This is the ***value function $V_\theta(t)$ over the 60-day trajectory***.

Inputs for Stage 4

- Weekly rewards $R_\theta(x_t)$ from S3
- Real temporal order of weeks
- Week-to-calendar/story phase mapping
- Discount factor γ (e.g., 0.99)

Step 1: Markov Reward Process

- We view each week t as a state, with immediate reward $R_\theta(x_t)$.
- Transitions are fixed by the story: state t always goes to t+1 (no actions here).
- This makes a 54-step Markov Reward Process over weeks.

Step 2: Compute Long-Term Value $V_\theta(t)$

For each week t, we define the discounted value:

$$V_\theta(t) = R_\theta(x_t) + \gamma R_\theta(x_{t+1}) + \gamma^2 R_\theta(x_{t+2}) + \dots$$

We compute this efficiently backwards from the end.

$V_\theta(t)$ represents "how valuable this week is, plus how valuable the rest of the story looks from here."

Step 3: Discover High/Low-Value Segments

- Rank all weeks by $V_\theta(t)$ from highest to lowest.
- Take the top-value starting weeks and map them back to concrete periods (e.g. "Valentine prep, days 1–7"; "late recovery, days 42–48").
- Take the lowest-value starting weeks and map them to their periods (e.g. "election peak, days 25–31").
  - **High-value:** Positive valence, structured work, supportive social, low conflict.
  - **Low-value:** Sustained conflict, election stress, emotional labour.

Outputs from Stage 4

- A value curve $V_\theta(t)$ for the 54 weeks.
- Automatically identified "high-value" and "low-value" life segments.
- Human-readable IRL discoveries on what drives good/bad long-term emotional trajectories, aligned with Big-5 personality.

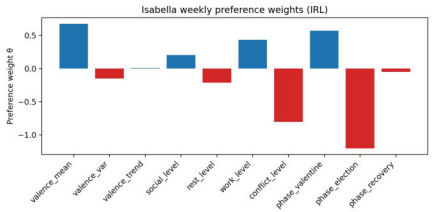
Temporal Dynamics: How θ Evolves

The true innovation: **θ is not static**. By learning θ over different periods, we can observe shifts in her preferences, reflecting personality growth through experience.

Big Five (Static)	IRL (60 Days)	Vθ(t) Trajectory
Self-report	Behavioral inference	Temporal dynamics
Isabella is conscientious / agreeable / extraverted."	"Isabella values structured work, low conflict, and positive weeks the most."	"Weeks 3–5 form a high-value turning point; election weeks mark a sustained low-value phase."
No explicit time axis.		

Experimental Results & Validation

1. Weekly preference IRL – learned θ

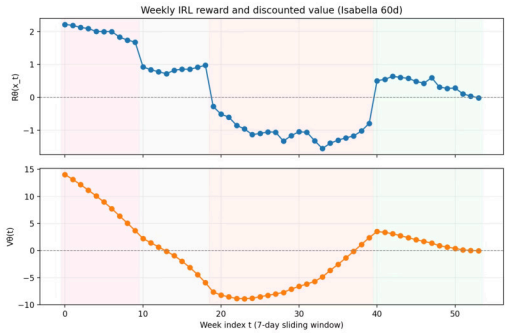


Using weekly proxy scores r_t and 10-D weekly states ϕ(s_t), we build weighted preferences between weeks and fit a linear reward $R_\theta(x_t) = \theta^T x_t$.

Interpretation:

- prefers positive, stable, structured-work, social weeks; strongly penalizes conflict-heavy election weeks.

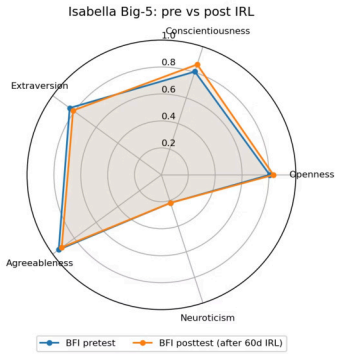
2. Trajectory value Vθ(t) and story alignment



What it shows:

- Top panel: weekly IRL reward $R_\theta(x_t)$ over 54 weeks.
- Bottom panel: discounted value $V_\theta(t)$.
- **Background shaded by phase (Valentine, election, recovery, other)**

3. Isabella Big-5: Pre vs Post 60-day IRL



What it shows:

- Overall shape is very stable → IRL does not distort her personality.
- Small, interpretable shifts

4. Big Five Validation: Before → After IRL

We assessed if the learned θ aligns with Isabella's independently measured Big Five personality profile.

Trait	Score Change	θ Evidence	Interpretation
Openness (O)	0.81 → 0.83	(Not reflected in θ)	Stable; comes from story context (art/community)
Conscientiousness (C)	0.81 → 0.86	work_level = +0.44 valence_var = −0.15	✓ Prefers structured work + stable emotions
Extraversion (E)	0.84 → 0.81	social_level = +0.21 phase_valentine = +0.57	✓ Values social connection, but not at any cost
Agreeableness (A)	0.94 → 0.92	conflict_level = −0.80 phase_election = −1.20	✓ Strongly averse to conflict; set boundaries under stress
Neuroticism (N)	0.22 → 0.22	valence_mean = +0.68 valence_var = −0.15	✓ Emotionally stable; prefers steady positivity over chaos

Finding: The IRL model never ***saw her Big Five scores,*** yet θ is psychologically coherent. This validates that we're capturing real personality-driven values, not noise.

Discussion: Limitations and Future Directions

Current Work Limitations

1. Data Scale

Only one individual, 60 days, 30 annotated preference pairs.

Next step: Expand to multiple participants to verify method generalisability.

2. Linear Preference Assumption

$R_\theta(s) = \theta^\top \phi(s)$ cannot capture nonlinear interactions between emotional features (e.g., "engagement + pressure" multiplicative effects).

Improvement: Explore nonlinear IRL (neural network reward functions).

3. Fixed Weekly Features

ϕ is hand-designed 6-dimensional features, potentially missing other important emotional dimensions. **Improvement**: Automatic feature learning (joint optimization of feature learning and IRL).

4. Missing Transition Model

We assume fixed transitions, unable to generate counterfactual reasoning like "if companion robot takes action A, how will emotions evolve".

Improvement: Learn transition model $P(s_{t+1}|s_t, a_t)$ to support policy optimization.

Deepening Scientific Significance

This work lays the foundation for a larger research question:

How can we design a companion agent in cyber-physical systems (CPS) that
respects users' long-term value systems rather than merely maximizing current satisfaction?

This question transcends the specific application of "emotional companionship", involving:

- **Ethical AI**: Robots should learn users' authentic preferences, not surface-level satisfaction
- **Human-Machine Value Alignment**: How to enable intelligent agents to understand human deep values, not just optimise metrics
- **Interpretable Reinforcement Learning**: θ parameter interpretability enables users to understand "why the robot accompanies me this way"

Summary

- Q1: Why Choose Reinforcement Learning (RL)?
  - Emotional companionship is fundamentally a sequential decision problem — states are multi-week emotional patterns, rewards are implicit in long-term preferences.
  - RL's MRP framework naturally represents this process, whilst IRL is the only method that can reverse-engineer reward functions from human behavioral preferences.
- Q2: Why Choose This Method?
  - We chose preference-based IRL (rather than direct RL) because our data takes the form of "which week is preferred" preference pairs, not explicit (s,r) annotations.
  - BiLSTM + attention state encoding balances interpretability and expressiveness.
  - MRP rather than MDP because we currently observe fixed trajectories — our goal is understanding, not control.
- Q3: What Are the Contributions?
  - Problem redefinition: From "maximizing single-instance happiness" to "respecting long-term preferences"
  - Complete technical loop: Fusion → encoding → IRL learning → value assessment
  - Counter-intuitive findings: Optimal emotion isn't "happiest" but "engaged + meaningful"
  - Foundation laying: Providing interpretable preference models for future emotional companionship policy learning