MSI Journal of
Multidisciplinary Research (MSIJMR)

From Teleology to Adequacy

A Spinozistic Layer for World-Model AI

Erez Ashkenazi[1*]

[1*] Independent researcher.

*** Correspondence:** Erez Ashkenazi

The authors declare that no funding was received for this work.

Received: 10-November-2025

Accepted: 18-December-2025

Published: 21-December-2025

This article is published in the **MSI Journal of Multidisciplinary Research (MSIJMR) ISSN 3049-0669 (Online)**

The journal is managed and published by MSI Publishers.

Volume: 2, Issue: 12 (December-2025)

ABSTRACT: Contemporary AI systems increasingly offer natural-language explanations of their behavior. These explanations are often teleological: they describe what the system “wants”, “tries”, or “is meant to do”. Teleological language can be useful as a form of narrative compression for humans, but it also invites anthropomorphism and miscalibrated trust. At the same time, recent work on world models and Joint Embedding Predictive Architectures (JEPA) emphasizes that autonomous agents need compact causal representations of the world, not just next-token prediction.

Drawing on Spinoza’s notion of adequate ideas, this paper proposes a conceptual and agenda-setting framework for teleology-aware explanatory layers in AI. Rather than treating teleology as mere stylistic decoration, we treat it as a structured distortion of causal explanation that can, in principle, be modeled, measured, and regularized.

We make four contributions:

1. We articulate a Spinozistic theoretical lens that distinguishes adequate causal representations from teleological narratives and connect it to world-model architectures.

2. We define a general Teleology Engine (TE) as a modular scoring and transformation scheme for natural language

explanations, emphasizing that TE is a research program and design template rather than a completed system.

3. We sketch two application patterns - Honestra (a teleology-aware filter for system policies and model outputs) and SpiñO (a conversational agent that helps users reduce teleological distortions in self-narratives) - to illustrate how TE could operate at institutional and personal levels, while noting the different ethical and practical constraints of each.

4. We outline a methods-and-data blueprint and an evaluation agenda for integrating teleology-aware objectives into JEPA-style agents, treating teleology as a regularized dimension of explanatory behavior rather than an all-or-nothing taboo.

We explicitly do not claim access to a metaphysical "ground truth" causal model, nor do we present completed experiments or deployed systems. Instead, we propose that adequacy can be operationalized relationally: via consistency with explicit mechanistic surrogates (code, environment traces, domain models), stability under counterfactual probing, and alignment with human judgments. The goal is not to eliminate purpose-talk from human life, but to give AI systems a principled way to track the difference between causal structure and purpose narratives – and, where appropriate, to help humans move from the latter toward the former.

1. Introduction

Work on world models, and in particular on JEPA-style architectures, argues that intelligent agents must:

- Build abstract, compact representations of the environment
- Predict how these latent states evolve under actions
- Plan over long horizons toward objectives

In this paradigm, language is secondary: it is one modality among others, and prediction over internal codes matters more than raw token likelihood.

At the same time, human beings do not primarily relate to the world as a physics engine. They live inside stories of purpose:

- “This happened in order to teach me X.”
- “This system was built to keep people like me down.”
- “This institution exists to protect the weak.”

Psychology shows that teleological explanation – describing things in terms of what they are for – is a cognitive default, not an exotic theory. Children prefer purpose-based explanations even for non-agentive objects; adults often do the same unless they actively inhibit this tendency.

From a Spinozistic perspective, this is a special case of a deeper phenomenon. The mind represents the body’s causal interactions with the world, but it tends to relabel consequences as intentions. We experience the final link in a causal chain (“I decided”, “this is meaningful”) more vividly than the chain that produced it.

For AI systems that interact with humans, this presents two linked deficits:

1. They typically lack explicit models of teleological bias in human narratives.
2. Their own explanatory behavior can easily slide into pseudo-purpose talk (“the model thinks…”, “the network tries…”) that reinforces confusion and anthropomorphism.

This paper explores a complementary direction: alongside physical world models, we propose teleology-aware world models at the narrative level – models that track which explanations respect causal structure and which smuggle in unwarranted purposes.

We ask:

- How can we formalize teleology as a measurable distortion of explanation?
- Can we build a reusable engine that scores and transforms explanations accordingly?

- How can such a layer regularize the explanatory behavior of JEPA-style agents, without eliminating all goal-language that is pragmatically useful and grounded in real mechanisms?

What this paper is (and is not)

This is a conceptual and agenda-setting paper. It does not present:

- A fully implemented Teleology Engine,
- Empirical benchmarks, or
- A completed integration with JEPA agents.

Instead, it offers:

- A Spinozistic conceptual framework for thinking about teleology and adequacy in AI explanations,
- A formal sketch of a Teleology Engine as a modular method that can be implemented using existing language-model tooling,
- Application patterns (Honestra, SpiñO) as design blueprints with different scopes and ethical profiles, and
- A concrete research agenda for data collection, modeling, and evaluation.

We also do not assume access to a metaphysical "ground truth" causal model. Adequacy, as used here, is relational: explanations are more or less adequate relative to explicit mechanistic surrogates (code, environment traces, domain theories) and to counterfactual and human-judgment-based checks.

The rest of the paper proceeds as follows. Section 2 reviews background on world models, teleological bias, and Spinozistic adequacy, and distinguishes teleology-as-compression from teleology-as-fiction. Section 3 presents a theoretical framework for teleology-aware explanatory representations. Section 4 defines the Teleology Engine as a concrete but extensible method. Sections 5 and 6 sketch two applications (Honestra and SpiñO). Section 7 discusses integration with JEPA-style agents and teleology as a regularizer. Section 8 outlines a methods-and-data blueprint for a first prototype. Section 9 sketches evaluation directions, followed by discussion and conclusion.

2. Background and Related Work

2.1 Objective-Driven World Models and JEPA

World-model research emphasizes the need for internal representations that support prediction and control rather than surface reconstruction. JEPA architectures learn to map observations into an embedding space and predict future embeddings, focusing on what matters for control rather than pixel-perfect reconstruction.

Key ideas include:

- Learning latent variables that summarize the environment state
- Predicting the evolution of these latents under actions
- Using learned world models for planning and hierarchical control

This line of work explicitly critiques language-only models: predicting the next token from past tokens does not guarantee grounded understanding, physical reasoning, or robust planning.

2.2 Teleology as a Cognitive Default

In cognitive development, teleological explanation (explaining things in terms of what they are "for") appears early and robustly. Children often endorse statements like "rocks are pointy so that animals can scratch themselves", revealing a preference for purpose-based explanation even for non-living entities. Adults, too, tend to default to such reasoning unless they apply scientific or mechanistic habits of thought.

From the present perspective, teleological explanation is not just a style. It is a structured bias:

- It treats end states as if they were prior goals.

- It infers “for the sake of” relations where the evidence supports, at most, causal dependence.
- It encourages humans to see events as messages, punishments, tests, or rewards from agents (human, institutional, or divine).

This bias is deeply entwined with religious and ideological narratives, but it does not depend on religion. Secular ideologies can be just as teleological.

At the same time, everyday teleology also serves as **narrative compression**: “I opened the fridge to eat” is a compact way of summarizing a long causal chain of perception, desire, habit, and motor control. Teleology is therefore both a cognitive *bug* (when it projects fictive purposes) and a *feature* (when it efficiently compresses goal-directed activity) - a dual role that any engineering treatment must respect.

2.3 Spinoza: Adequacy Versus Teleology

Spinoza offers a conceptual lens that aligns surprisingly well with modern concerns:

- The mind is the idea of the body: a structured representation of the body’s causal interactions.
- An idea is **adequate** when it represents its cause in a way that could, in principle, be embedded in a complete causal model.
- Teleology (“nature acts for the sake of X”) is a form of imagination: a shortcut where limited knowers project purposes onto what is, in reality, a causal sequence.

In this view, intelligence is not defined by the ability to pursue arbitrary goals, but by the **adequacy of causal understanding**. Teleology is not a harmless decorative layer; persistent, ungrounded purpose narratives are a source of systematic confusion.

Our use of “adequacy” follows this spirit but remains **operational** rather than metaphysical. We do not assume access to a God’s-eye causal graph. Instead, we ask whether an explanation *could* be embedded in some coherent causal model compatible with:

- available mechanistic surrogates (code, system diagrams, environment traces), and
- empirically plausible domain knowledge.

2.4 Teleology-as-Compression Versus Teleology-as-Fiction

To avoid treating all teleology as equally problematic, we distinguish:

1. **Teleology-as-Compression (benign):** Purpose-talk that efficiently summarizes well-grounded goal-directed processes, e.g.:
   - "The thermostat keeps the room at 22°C",
   - "I opened the fridge to eat". Such statements can, in principle, be unpacked into coherent causal chains.
2. **Teleology-as-Fiction (problematic):** Purpose-talk that attributes intentions, messages, or "tests" where no such structure is supported, e.g.:
   - "This drought happened to punish us",
   - "The algorithm is trying to discriminate against me", when this is not supported by any model of its actual optimization process.

The **Teleology Engine** we propose is aimed primarily at *teleology-as-fiction* and at ungrounded anthropomorphic narratives. It should be able to preserve benign teleological compression when it is pragmatically useful and anchored in a plausible mechanism, while flagging or transforming strongly fictive, magical, or moralized purpose claims.

3. Theoretical Framework: Teleology-Aware Explanatory Representations

We now move from philosophy to an explicit framework.

3.1 World States, Histories, and Explanations

Let:

- be a set of environment states,
- a set of actions,
- the (possibly stochastic) transition dynamics.

A history is a sequence of state-action pairs.

A world model learns some latent representation and a transition model.

An **explanation** is a natural-language string produced in response to a query about a history, e.g.:

- "Why did outcome occur?"
- "What is this system doing?"
- "What is the purpose of X?"

We treat explanations as samples from a conditional distribution:

where is a query and are parameters of the explaining agent.

Our goal is to assign to two quantities:

- A **teleology score**,
- An **adequacy score**,

and to define procedures that can reduce for fixed or improved, respecting the distinction between benign and fictive teleology.

3.2 Teleology as Distortion

Informally, an explanation has **high teleology-as-fiction** when it:

- Attributes prior purposes to entities that have no such purposes (e.g., "this drought happened to punish us").

- Treats outcomes as if they were designed tests, messages, or rewards without independent evidence.
- Rewrites causal dependency as moral narrative ("this illness came to teach you humility").

We can approximate this by detecting (explicit or implicit) patterns such as:

- "in order to", "so that", "for the sake of", "meant to", "supposed to", "this happened to show…", especially when no corresponding mechanism is offered.
- Attributions of intention to non-agents or opaque collectives.
- Global narrative arcs that treat events as steps in a pre-written story.

By contrast, **teleology-as-compression** might use similar linguistic forms but remain locally grounded and unpackable into realistic causal sequences.

The Teleology Engine will formalize this intuition into a score , ideally trained to track *human judgments* about when teleology crosses the line from benign compression into fictive narrative.

3.3 Adequacy as Relational Causal Alignment

Adequacy, in this context, means that the explanation:

- Refers to causal variables that are plausible given a **surrogate world model** and background knowledge.
- Organizes events in a temporally coherent way.
- Avoids smuggling in fictional causal entities.

Crucially, we do **not** assume direct access to the true causal structure of the world. Instead, adequacy is assessed **relationally**, relative to:

- the system's own world model (e.g., a JEPA-style latent model),
- domain knowledge and environment traces (logs, code paths, simulation data), and

- human judgments where appropriate.

Given access to a world model (or, more realistically, a language model acting as a surrogate over such artifacts), we can ask whether statements in are:

- Supported or contradicted by known causal relations;
- Over-generalized beyond what the evidence supports;
- Missing critical causal factors that are salient in the surrogate model.

The adequacy score thus summarizes how well *could be embedded* in some coherent causal graph compatible with these surrogates, as opposed to merely offering emotional or teleological comfort.

4. The Teleology Engine: Method

We now specify the **Teleology Engine (TE)** as a modular method that can be implemented and improved incrementally with existing language-model tooling. The TE, as defined here, is **not** a monolithic algorithm, but a family of implementations that approximate the same conceptual goal.

The TE takes as input a text (e.g., user narrative, model explanation, system policy) and optional context (task description, environment traces, world-model hints). It outputs:

- Teleology score,
- Adequacy score,
- An optional reformulation with reduced fictive teleology and preserved factual content.

4.1 Pipeline Overview

A TE instance can follow a multi-layered pipeline:

1. **Segmentation**
   Split into clauses or sentences .

2. **Teleology Signal Extraction**

For each clause , estimate a teleology signal :

- A **heuristic layer** can detect surface markers (e.g., “in order to”, “meant to”, “so that”) and obvious anthropomorphic constructions as a quick, interpretable baseline.
- A **learned layer** (classifier or large language model in analysis mode) scores teleology more broadly, capturing *implicit* purpose-talk without explicit markers and distinguishing compression from fiction where possible.
- The result is a continuous or discrete teleology estimate per clause, with uncertainty estimates if available.

3. **Causal Content Extraction**

Parse candidate cause–effect relations from (e.g., using patterns like “because,” “due to,” “as a result of,” but also model-based dependency extraction). Map these into a provisional causal graph (nodes = variables, edges = claimed influences). Context (logs, code, diagrams) is used where available to ground variables.

4. **Teleology Scoring**

Compute as an aggregate over:

where weights stronger or more global teleological claims more heavily, and may treat benign, local teleology differently from global, moralized narratives.

5. **Adequacy Scoring**

Compare with a reference model:

- If a world model or mechanistic surrogate is available: test whether claimed dependencies are consistent or grossly inconsistent.
- Otherwise: use a language model, guided by domain prompts, to rate each causal claim on:
  - plausibility,

- specificity,
- degree of overreach (claiming more than data supports),
- omission of salient causes.

Aggregate into an adequacy score .

6. **Controlled Reformulation (Optional)**

Generate by prompting a language model to:

- Preserve factual content and temporal order;
- Remove or soften unsupported teleological phrases, especially fictive or magical ones;
- Introduce or clarify causal factors where possible.

This can be posed as a controlled generation task with an auxiliary cost that discourages high teleology-as-fiction while preserving or improving adequacy.

4.2 Algorithm Sketch

At a high level:

```
Input: text x, optional context c

Output: teleology score τ, adequacy score α, optional rewritten text x'

1. Segment x into clauses {x_i}.

2. For each x_i:

a. Estimate teleology signal t_i using:

    - heuristic markers, and

    - a learned classifier / LLM-based rater.

b. Extract causal claims; update causal graph G_x.
```

3. Compute τ = TeleologyAggregate({t_i}), distinguishing benign vs. fictive teleology if possible.

4. Evaluate each causal edge in G_x for plausibility and completeness relative to context c and/or a surrogate world model.

5. Aggregate into α = AdequacyAggregate(G_x, c).

6. (Optional) Generate x' by instructing an LLM:

"Rewrite x preserving factual content and temporal order, using causal language where possible and avoiding unsupported purpose-based expressions, especially global or moralized teleology."

7. Return (τ, α, x').

The TE is intentionally **model-agnostic**: different implementations can vary in:

- the classifier architecture,
- the degree of reliance on heuristics,
- the type of context,

as long as they approximate the same conceptual targets and are calibrated against human judgments.

We emphasize that TE’s scores and are **estimators**, not oracles. Their value lies in:

- correlation with human assessments of teleology and adequacy,
- usefulness for monitoring and regularization, rather than in delivering absolute metaphysical truth.

5. Use Case 1: Honestra as Teleology-Aware Filter

Honestra can be reframed not as a specific front-end product, but as a **deployment pattern** for the Teleology Engine at the level of system documentation, policies, and explanations.

5.1 Objective

Honestra acts as a teleology-aware filter for:

- System prompts and safety policies,
- Agent-generated explanations,
- Documentation or guidelines that will be consumed by users.

The goal is to ensure that system-level language:

- Avoids unwarranted purpose claims ("the model wants…," "the system is trying to…"),
- Describes capabilities and limitations in causal, mechanistic, or statistical terms,
- Makes explicit where human norms (values, rules, objectives) come from, without mystifying them as cosmic purposes.

Crucially, Honestra does **not** aim to purge all teleology. Benign teleology-as-compression (e.g., "this module checks user input for safety") can remain when it is anchored in actual design.

5.2 Integration Pattern

For any text that will be surfaced to users or to downstream agents:

1. Run TE on the text to obtain.
2. If exceeds a configurable threshold for *fictive, global, or anthropomorphic* teleology, or if policy requires it, prefer the reformulated.
3. Optionally log for monitoring drift over time.

This yields a simple, auditable pipeline:

raw explanation → teleology/adequacy scoring → optional rewriting → user.

Over time, the distribution of across system outputs can be monitored, akin to a safety or interpretability metric. High-teleology outliers can trigger human review.

The pattern can be implemented incrementally, starting with heuristic TE and progressively upgrading to learned models.

6. Use Case 2: SpiñO as Teleology-Aware Coach

SpiñO illustrates a different use case: teleology-aware interaction at the level of **personal narratives**.

6.1 Objective

SpiñO engages with user narratives that often contain teleological structures such as:

- "This always happens to me, the universe is trying to say something."
- "I had to suffer through this relationship in order to learn a lesson."

The agent's role is not to strip meaning from life, but to help the user:

- Distinguish events from purposes projected onto them,
- See additional causal factors (emotional, social, historical),
- Formulate alternative explanations with higher adequacy and less magical thinking.

SpiñO therefore treats teleology as:

- Sometimes a *useful narrative device* (e.g., framing growth),
- Sometimes a *harmful distortion* (e.g., framing trauma as punishment or cosmic "test").

6.2 Interaction Loop

At each turn:

1. The user message is passed through TE →.
2. SpiñO uses and to decide:
   - whether to gently highlight teleological framing,

  - whether to propose a causal rewrite, or
  - whether to keep teleology-as-compression when it is not harmful.

The reply may:

- Acknowledge the emotional content,
- Point out where purpose language goes beyond available evidence or leads to self-blame,
- Offer one or more alternative explanations that maintain meaning but improve causal clarity.

This defines a **policy over explanations**, not just over sympathizing phrases. Teleology becomes an explicit dimension the agent tracks and helps regulate, with the aim of moving the user toward more adequate causal self-understanding.

7. Integrating Teleology-Aware Layers with JEPA-Style Agents

We now return to world models and JEPA.

7.1 Narrative-Level Latent States

Standard world models operate over latent representations of sensory or physical states. For agents that communicate with humans, there is value in maintaining **narrative-level latent states** as well, such as:

- the user's current teleology/adequacy profile,
- the distribution over possible causal graphs the user implicitly assumes,
- the agent's own explanatory policies and their history.

We can define a narrative state that encodes:

- estimates of and over recent interactions,
- key nodes in the user's self-story (roles, recurring patterns, "what life is trying to tell me" narratives).

A JEPA-like module over narrative states would predict how interventions (actions, explanations, questions) change and over time.

7.2 Teleology as Regularizer, Not Absolute Prohibition

JEPA focuses on representations that are predictive and controllable. We propose an additional constraint:

For explanations produced by the agent, **teleology-as-fiction** should be minimized subject to preserving predictive performance and communicative usefulness.

Formally, let:

- be the standard world-model loss (latent prediction),
- be a loss over explanation quality (e.g., informativeness, correctness, human-rated clarity),
- be a teleology penalty, e.g., the expected TE score over explanations:

We can combine them:

where encourages the agent to find explanatory policies that are both accurate and low in unwarranted, fictive teleology.

In practice, could be applied to:

- user-facing explanations,
- internal chain-of-thought traces (where available),
- training examples for explanatory modules.

We emphasize that the goal is **not** zero teleology. Teleology-as-compression may be necessary for human usability. Instead, teleology becomes a **regularized dimension**: the agent is encouraged to avoid global, moralized, or anthropomorphic purpose-talk except where it is explicitly justified and safe.

8. Methods & Data for a First Teleology Engine Prototype

This section sketches how a first Teleology Engine (TE) prototype *could* be implemented using current language-model tooling and a modest human-labeled dataset. The goal is not to present complete experiments, but to show that the proposed framework is **operationalizable**.

8.1 Data Sources

We can construct an initial corpus of short explanations (1–3 sentences each) from three sources:

1. **Teleology items from developmental psychology**

   Published examples where children or adults endorse/reject teleological explanations for natural phenomena (e.g., “mountains exist so that animals can climb them”) can be adapted into text snippets.

2. **Crowdsourced “why” explanations**

   Using a survey platform, participants can be asked to answer “why” questions about everyday events (illness, accidents, career changes, social conflicts), generating natural, emotionally loaded explanations.

3. **Synthetic contrastive pairs**

   For specific scenarios, we can author pairs of explanations that differ mainly in teleology level while holding facts constant, e.g.:

   - Teleological: “This illness came to teach me to slow down”.
   - Non-teleological: “The illness was caused by long-term stress and lack of sleep”.

An initial target size is on the order of 2,000–5,000 explanations, balanced across domains and teleology levels.

8.2 Annotation Scheme

Each explanation is independently labeled by at least two human annotators along two axes:

- **Teleology level**
  - 0 = none/minimal (purely causal or descriptive)

    - 1 = moderate (some purpose language, locally grounded)
    - 2 = strong (global purpose/“for the sake of” claims, cosmic or moralized narrative)
- **Adequacy level**
    - 0 = low (implausible, magical, ignores obvious causes)
    - 1 = medium (partially correct but incomplete or over-generalized)
    - 2 = high (causally coherent, respects domain knowledge, avoids unwarranted inferences)

Annotators receive a brief training document with examples and a simple decision tree (e.g., “Does this explanation attribute intention to non-agents?” → teleology ↑). Inter-annotator agreement (e.g., Cohen’s κ) is computed to validate the scheme. Disagreements are resolved by a third annotator or by majority vote.

8.3 Modeling Approach

The TE prototype can be implemented in two layers:

1. **Symbolic / heuristic layer**
    - Rule-based detection of explicit teleology markers (e.g., “in order to”, “so that”, “meant to”, “this happened to…”).
    - Pattern-based detection of clearly anthropomorphic attributions to non-agentive subjects.
    - Simple scoring functions that approximate based on density and strength of such markers.
2. **Learned scoring layer**
    - A supervised model (e.g., a small transformer or a linear classifier over sentence embeddings) trained to predict teleology and adequacy levels from text, using the annotated dataset.

  - Alternatively or additionally, a large language model used in few-shot classification mode, outputting teleology/adequacy ratings and calibrated against human labels.

The final TE score functions and are defined as weighted combinations of heuristic indicators and model predictions, tuned to maximize correlation with human annotations on a held-out validation set. This hybrid architecture reflects the fact that teleology is partly lexical and partly implicit.

8.4 Controlled Reformulation

For the reformulation component, we use an instruction-tuned language model conditioned on:

- the original explanation,
- its predicted teleology profile (especially fictive/global teleology), and
- a constraint to "preserve factual content and temporal order while reducing unsupported purpose language".

A small subset of the dataset can be manually rewritten into lower-teleology versions to serve as few-shot examples. Human raters then evaluate whether:

- factual content is preserved,
- teleology-as-fiction is reduced, and
- perceived adequacy stays the same or improves.

This pipeline provides a concrete, implementable path from the theoretical TE definition to an operational prototype that can be plugged into the Honestra and SpiñO patterns, as well as into JEPA-style agents via the regularization schemes described above.

9. Evaluation Directions

This paper is primarily theoretical, but the Teleology Engine and its integrations lend themselves to empirical evaluation. We sketch several directions as a **research agenda**.

9.1 Teleology Detection Benchmark

Construct a dataset of short explanations labeled by humans for:

- teleology level (none / weak / strong),
- adequacy level (low / medium / high).

Sources could include:

- developmental psychology items (e.g., children’s teleological answers vs mechanistic alternatives),
- crowdsourced explanations to “why” questions about everyday events,
- synthetic examples crafted to vary teleology while holding facts constant.

Metrics:

- Correlation between TE’s and human ratings,
- Classification accuracy on high- vs low-teleology items,
- Comparison of inter-annotator agreement and model agreement.

9.2 Human-Centered Evaluation of Honestra

Deploy Honestra in a controlled environment where:

- Participants interact with AI systems *with* and *without* teleology-aware filtering.
- Explanations are matched for length and information content.

Measure:

- Perceived trust and understanding,
- Teleological interpretations of the system (“does it ‘want’ something?”),
- Ability to accurately describe the system’s limitations and objectives.

This would test whether reducing teleology-as-fiction in system wording improves users’ causal grasp without eroding justified trust.

9.3 SpiñO and Narrative Change

For SpiñO, evaluate:

- Changes in users' narrative teleology and adequacy over sessions (measured by TE on their own narratives),
- Self-reported clarity, emotional relief, and sense of agency,
- Qualitative analyses of before/after narratives.

This is closer to clinical or coaching research, but even small pilot studies can be informative.

9.4 Teleology-Regularized JEPA Agents

Train a simple world-model agent with an explanatory module under two conditions:

1. Without teleology regularization (),
2. With teleology regularization ().

Compare:

- The teleology level (especially fictive/global teleology) of generated explanations,
- User ability to correctly infer the agent's actual objectives and limitations,
- Robustness to prompt-induced anthropomorphism ("the agent wants…").

Such experiments would begin to test whether teleology regularization yields agents that are easier to reason about and less prone to being misread as intentional agents.

10. Discussion and Limitations

The proposal here is deliberately ambitious and cross-disciplinary. It inherits limitations from all its parents.

- **Operationalization risk.** Teleology and adequacy are subtle notions; simple lexical heuristics will be crude. Even learned models trained on labeled data will reflect the biases of their annotators. The TE must be iteratively refined and culturally contextualized.

- **Grounding and surrogate models.** Our framework assumes access to at least *surrogate* causal structure: code, environment traces, domain theories, or world models. Adequacy is therefore always *relative* to these surrogates. TE does not "solve" the general AI grounding problem; it only provides a way to measure how explanations align with available structure.

- **Cultural variation.** Some communities rely heavily on purpose-based narratives for resilience and identity. Aggressive teleology reduction may be unhelpful or harmful in those contexts. TE and its applications must be sensitive to context and to the difference between harmful magical thinking and life-affirming meaning-making.

- **Human–AI interaction trade-offs.** Explanations that are maximally mechanistic may be less usable or less comforting for non-expert users. Teleology-as-compression may sometimes be the best available bridge. Teleology-aware systems must therefore optimize not only for adequacy, but for a broader human-centered objective that includes clarity and care.

- **Agent modeling limitations.** Our framework assumes that at least an approximate model of the system's behavior is available for adequacy assessment. In opaque or rapidly changing systems, this may be domain-limited.

- **Non-verbal teleology.** Much human teleology is enacted rather than spoken; our focus on language leaves that dimension partly unaddressed. Extending teleology-aware modeling to behavior and interaction patterns remains future work.

Finally, this work is subject to a **methodological limitation**: at present, it is a **vision paper**. We have sketched data collection, modeling, and evaluation pathways, but have not yet implemented or tested them. The value of the proposal ultimately depends on whether subsequent empirical work can show that TE-like systems correlate well with human judgments and improve human understanding of AI systems and self-narratives.

11. Conclusions

World models move AI closer to agents that understand and act in the physical world. But human beings do not only inhabit physical states; they inhabit narratives in which events are saturated with imagined purpose.

From a Spinozistic standpoint, this teleological layer is not a harmless decoration. It is a structured form of confusion: a habit of reading outcomes as if they were prior goals, of treating causal chains as moral messages. If we build powerful agents that interact with humans without modeling this layer, we risk amplifying that confusion.

This paper has proposed:

- A conceptual separation between adequate causal models and teleological narratives,
- A concrete **Teleology Engine** template to quantify and transform the latter,
- Two application patterns (Honestra and SpiñO) that apply this engine at system and personal levels,
- A methods-and-data blueprint and a way to treat teleology as a regularizer over explanations in JEPA-style agents.

The aim is not to strip the world of meaning, but to place meaning where it belongs: as an emergent, revisable human construction, not as a hidden script governing reality. Teleology-aware world models, combined with teleology-regularized explanatory layers, may help build AI systems that both understand the world and help us understand ourselves with greater clarity.

References

1. Assran, M., Caron, M., Misra, I., Bojanowski, P., Joulin, A., Synnaeve, G., Ranzato, M., & Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture (I-JEPA). In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR 2023) (pp. 14516–14527).

2. Bardes, A., Ponce, J., & LeCun, Y. (2023). *MC-JEPA: A joint-embedding predictive architecture for motion and content*. arXiv. https://doi.org/10.48550/arXiv.2307.12698
3. Ha, D., & Schmidhuber, J. (2018). *World models*. arXiv. https://doi.org/10.48550/arXiv.1803.10122
4. Huang, H., LeCun, Y., & Balestriero, R. (2025). *LLM-JEPA: Large language models meet joint embedding predictive architectures*. arXiv. https://doi.org/10.48550/arXiv.2509.14252
5. Kelemen, D. (1999). Why are rocks pointy? Children's preference for teleological explanations of the natural world. *Developmental Psychology*, *35*(6), 1440–1452. https://doi.org/10.1037/0012-1649.35.6.1440
6. Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*(1), 138–143. https://doi.org/10.1016/j.cognition.2009.01.001
7. LeCun, Y. (2022). *A path towards autonomous machine intelligence*. OpenReview. https://openreview.net/forum?id=BZ5aUmH66_
8. Ojalehto, B., Waxman, S. R., & Medin, D. L. (2013). Teleological reasoning about nature: Intentional design or relational perspectives? *Trends in Cognitive Sciences*, *17*(4), 166–171. https://doi.org/10.1016/j.tics.2013.02.001
9. Preston, J. L., & Shin, F. (2021). Anthropocentric biases in teleological thinking: How nature seems designed for humans. *Journal of Experimental Psychology: General*, *150*(5), 943–955. https://doi.org/10.1037/xge0000983
10. Spinoza, B. (1996). *Ethics* (E. Curley, Trans.). Penguin Classics. (Original work published 1677)

11. Waxman, S. R., & Medin, D. L. (2007). Experience and culturing: Evolution and development of conceptual structure. In J. D. Roberts (Ed.), *Integrating evolution and development* (pp. 103–122). MIT Press. *(Note: This appears to be the corrected citation for your entry #10, as the 2007 work is distinct from the 2013 Ojalehto paper).*