UNWIND — MSSE Capstone Project

Michael Harris

Project Overview

UNWIND is a constrained AI system designed to address a specific failure mode common in conversational systems: **narrative reinforcement of distress**.

Rather than generating explanations, reassurance, or advice, UNWIND operates as a **training environment** that continuously redirects attention back to direct bodily sensation. The system is intentionally limited in scope to preserve behavioral integrity.

This repository contains the **Phase One implementation**, which validates the core mechanism through deterministic logic, automated tests, and scenario-based evaluation.

Problem Statement

Most AI systems intended for well-being drift toward:

- Reassurance
- Interpretation
- Narrative elaboration
- Emotional validation that reinforces story

While these behaviors feel supportive, they often **amplify identification with narrative**, increasing dependence and cognitive looping.

UNWIND was designed to test a different hypothesis:

That sustained attention to direct bodily sensation, without narrative engagement, reduces unnecessary suffering more reliably than explanation or reassurance.

Design Goals

Phase One focused on **mechanism integrity**, not warmth or personalization.

Key goals:

- Prevent narrative co-authoring
- Avoid emotional interpretation
- Enforce redirection to sensation
- Remain predictable and testable
- Support short, real-world use

The system prioritizes **constraint over flexibility** to avoid drift.

System Architecture

UNWIND is implemented as a **Python-based state machine** with strict control logic.

High-level components:

- **Classifier Layer**
  Identifies when user input shifts into:
  - Story
  - Explanation
  - Reassurance-seeking
  - Management or control attempts
- **Constraint Engine**
  Determines whether the system may respond directly or must redirect.
- **Redirection Logic**
  Returns attention to present-moment bodily sensation without interpretation.
- **Test Harness**
  Automated tests validate that narrative paths are consistently blocked.

The system does not generate open-ended therapeutic dialogue.

Repository Structure

```
unwind_v2/
├── controller/        # Core state machine and control logic
├── language/          # Classification and pattern detection
├── tests/             # Automated tests (pytest)
├── requirements.txt   # Python dependencies
├── pytest.ini         # Test configuration
├── CHANGELOG.md
├── scenario_output.txt
└── scenario_output2.txt
```

Supporting documentation is provided separately.

Running the Application

Requirements

- Python 3.9+
- pip

Install dependencies

```
pip install -r requirements.txt
```

Run the app

```
python main.py
```

(The app runs locally as a text-based interaction.)

Running Tests

Automated tests validate classifier behavior and state transitions.

```
pytest
```

Test coverage focuses on:

- Correct identification of narrative moves
- Reliable redirection behavior
- Prevention of conversational drift

Evaluation Approach

Rather than measuring emotional outcomes, Phase One evaluates:

- Behavioral consistency
- Constraint enforcement
- Drift prevention
- Predictability under repeated use

Scenario transcripts are included to demonstrate real interaction patterns and system responses.

Phase Two (Conceptual Only)

Phase Two is **not implemented** in this repository.

It proposes an **agentic attunement layer** that introduces limited acknowledgment while preserving constraint integrity.

Key design tension:

- Too rigid → alienating
- Too attuned → narrative reinforcement

Phase Two explores managing this tension through controlled deviation and continuous correction, similar to navigational systems.

Phase Two is documented conceptually only and is outside the scope of this capstone implementation.

Scope and Limitations

UNWIND is intentionally limited.

It is:

- Not therapy
- Not diagnostic
- Not a wellness chatbot
- Not a general conversational agent

It does not:

- Interpret emotions
- Provide advice
- Offer reassurance
- Attempt to resolve problems

Its purpose is to test whether **contact without narrative** can be trained through constraint.

Author

Michael Harris
MBA, Executive Background
MSSE Candidate (in progress)

Student ID Q160643471905991804

UNWIND emerged from both professional experience in systems design and a multi-year personal investigation into the mechanics of suffering and attention.