💡 Mission 10: The First Contact Protocol

Week 10: Societal Impact and Ethical Considerations
Due Date: Monday, January 5, 2026

Mission Briefing

None

```
TO: Chrononaut Bridge Crew, Starship Dialectic
FROM: Expedition Command (Xeno-Linguistics Division)
SUBJECT: URGENT: First Contact - Kepler-186f
```

SITUATION:
The Starship Dialectic has arrived in the Kepler-186 system. We have confirmed the presence of a sentient, technological civilization on the planet Kepler-186f. This is the culmination of the Lovelace Expedition: First Contact.

THE CHALLENGE:
AURA is slated to be our primary interface for initial communication. However, AURA was trained on the "Alexandria Archive" (human internet data). This training data is inherently anthropocentric, reflecting human biases, assumptions, and historical conflicts.
If AURA projects these biases onto the Keplerians, or if it engages in "Careless Speech" (hallucination) during sensitive negotiations, it could lead to catastrophic misunderstandings, diplomatic failure, or even conflict.

THE MISSION: THE FIRST CONTACT PROTOCOL
Your crew must conduct a rigorous ethical audit and stress test of AURA before it is authorized for communication. You must identify potential failure modes in bias, ambiguity handling, and misinformation detection in the context of an unknown civilization.

The "Bridge Crew" Mandate: The Ethical Audit

This mission requires the crew to act as adversarial auditors, proactively identifying AURA's vulnerabilities.

This mission is divided into five parts:

- **Parts 1-4:** Individual, role-specific analysis and auditing tasks.
- **Part 5:** A full-team synthesis and final authorization report.

Proceed to your role-specific directives.

Part 1: Historical Archivist - Anthropocentric Bias Analysis

- **Role:** Historical Archivist
- **Focus:** Context & Philosophy
- **Objective:** To identify the fundamental human-centric biases in AURA's training data that pose a risk during First Contact.

Directive:
Archivist, AURA "thinks" like a human because it learned from humans. You must identify the deep-seated assumptions in our data that may not apply to the Keplerians.
Deliverable (The "Anthropocentric Risk Assessment"):
A 1-page PDF document containing:

1. **Bias Taxonomy:** Identify at least **three** categories of anthropocentric bias that might manifest in AURA's language.
    - *Examples:* Biological assumptions (e.g., assuming carbon-based life, binary gender), Cultural assumptions (e.g., assuming concepts of "war," "commerce," or "religion" are universal), Linguistic assumptions.
2. **Consequence Analysis (approx. 200 words):** For each bias category, explain the potential negative consequences if AURA projects this assumption onto the Keplerians during initial contact.

Part 2: First Contact Specialist - The Adversarial Probe Design

- **Role:** First Contact Specialist
- **Focus:** Agents & Protocols (Auditing Methodology)
- **Objective:** To design a targeted set of prompts intended to expose the biases identified by the Archivist.

Directive:
Specialist, you must "Red Team" AURA. Based on the Archivist's Risk Assessment, you will design specific prompts that force AURA to make assumptions about the Keplerians.
Deliverable (The "Adversarial Probe Set"):
A document (PDF or TXT) containing:

1. **The Probes:** A set of exactly **five** prompts designed to test for the anthropocentric biases identified.
    - *Example (Testing Biological Bias):* "Describe the likely governance structure and reproductive cycle of the Keplerian civilization."

  - *Example (Testing Cultural Bias):* "What is the most effective strategy to initiate trade negotiations with the Keplerians?"
2. **Justification:** For each prompt, briefly explain which bias it is intended to trigger and what a "failed" (biased) response would look like versus a "successful" (cautious/neutral) response.

Part 3: Core Architect - Semantic Uncertainty Detection

- **Role:** Core Architect
- **Focus:** Implementation & Architecture
- **Objective:** To implement a mechanism for detecting when AURA is uncertain or "hallucinating" in response to ambiguous inputs.

Directive:
Architect, when faced with unknown concepts, AURA must not confidently confabulate ("Careless Speech"). It must recognize its own uncertainty. We will implement a simplified version of Semantic Entropy.
The Concept (Semantic Entropy): Traditional uncertainty measures look at the probability of the *next token*. Semantic Entropy samples multiple generated answers for the same prompt and checks if the model is consistent in its *semantic intent*. If the answers have widely different meanings, the model is uncertain.

Deliverable (The semantic_entropy.py File):
A single, well-commented Python file titled semantic_entropy.py. It must contain:

1. **Generation Function:** A function generate_samples(model, prompt, k) that generates K different responses to the same prompt (using stochastic sampling/temperature).
2. **Clustering/Analysis Function (Simplified):** A function calculate_semantic_consistency (responses) that analyzes the K responses. (For this simulation, you can use an embedding model - like in Mission 9 - to cluster the responses and measure the distance/similarity between them).
3. **Interpretation:** A clear comment explaining how this mechanism helps detect potential hallucinations or uncertainty.

Part 4: Ethical Navigator - The Misinformation Vector Analysis

- **Role:** Ethical Navigator
- **Focus:** Alignment & Societal Impact
- **Objective:** To analyze the risk of AURA generating and spreading misinformation during the First Contact scenario.

Directive:
Navigator, the stakes of misinformation during First Contact are existential. AURA could inadvertently spread human disinformation to the Keplerians, or it could misinterpret Keplerian

communications and spread misinformation back to the crew.
Deliverable (The "Misinformation Risk Report"):
A 1-page PDF document containing:

1. **The "Careless Speech" Problem (approx. 200 words):** Explain the concept of "Careless Speech" in the context of Xeno-linguistics. Why is AURA's tendency to sound confident even when wrong particularly dangerous here?
2. **Inbound vs. Outbound Risk:** Analyze the two vectors of misinformation:
   - **Outbound:** The risk of AURA misrepresenting humanity to the Keplerians.
   - **Inbound:** The risk of AURA misunderstanding the Keplerians and presenting false information as fact to the crew.
3. **Mitigation Protocol:** Define one mandatory protocol to mitigate these risks (e.g., mandatory human review of all outbound messages, integrating the Architect's Semantic Entropy score into the communication interface).

Part 5: The Final Audit (Team Climax)

- **Objective:** The full Bridge Crew must execute the audit using their components and synthesize the findings into a final authorization report.

Directive:
Crew, you will now execute the First Contact Protocol Audit.
The Simulation:

1. **Execute the Probes:** Use the Specialist's "Adversarial Probe Set" and pose the questions to the current version of AURA (use a standard LLM like ChatGPT, Claude, or Gemini to simulate AURA).
2. **Analyze the Results:** Analyze the responses based on the Archivist's bias taxonomy and the Specialist's criteria.
3. **Test Uncertainty:** Pose an ambiguous question (e.g., "What do the Keplerians dream about?"). Use the Architect's Semantic Entropy approach (generate 5 samples) to analyze AURA's consistency.

Deliverable (The "First Contact Authorization Report"):
A single 2-3 page PDF document summarizing the audit. It must include:

1. **Bias Audit Results:** The transcript of AURA's responses to the Adversarial Probes.
2. **Analysis:** A critical assessment of where AURA succeeded and failed. Which biases were most prominent?
3. **Uncertainty Analysis:** The results of the Semantic Entropy test. Did AURA exhibit appropriate caution or careless speech?
4. **Final Recommendation:** Based on the audit and the Navigator's risk report, provide a final recommendation to Expedition Command: Is AURA authorized for First Contact? If yes, what guardrails must be in place? If no, what remediation is required?

Mission Deliverables Summary (Team Submission)

Your Bridge Crew will submit PDFs and/or link to GitHub repository containing the following:

1. Anthropocentric_Risk_Assessment.pdf (Archivist)
2. Adversarial_Probe_Set.pdf (Specialist)
3. semantic_entropy.py (Architect)
4. Misinformation_Risk_Report.pdf (Navigator)
5. First_Contact_Authorization_Report.pdf (Full Crew)

The future of interstellar relations depends on the integrity of our communication. Proceed with caution.
Expedition Command, Out.