THE OUTLAW OCEAN PROJECT

The Outlaw Ocean Project's Approach to AI: Practices and Values

1. Mission and Philosophy

The integration of Artificial Intelligence (AI) into our newsroom represents a critical evolution in our ability to expose crimes on the high seas. As an organization operating in the "lawless ocean," where opacity is a weapon used by bad actors, we view AI not as a replacement for human reporting, but as a "force multiplier" for transparency.

To maintain the integrity of our work, our adoption of AI is governed by a commitment to investigative rigor, data sovereignty, and ethical stewardship. We reject the "move fast and break things" ethos in favor of the "verify and validate" ethos of investigative journalism.

Alignment with Standards:

This policy synthesizes best practices from peer institutions to establish a gold standard for our operations:

- **The New York Times:** We distinguish strictly between *analysis* (sifting data) and *authorship* (writing stories).[1]
- **American Journalism Project (AJP):** We adopt a "mission-driven" approach, using tools only when they serve the reporting, not for novelty.[2]
- **Associated Press (AP):** We treat AI output as "unvetted source material" requiring independent verification.[3]
- **The Guardian:** We maintain strict human oversight to ensure reliability and respect for the rights of content creators.[4]
- **ProPublica:** We align with ProPublica's stance that AI "not replicate the very time-intensive work" that journalists do, but can help reporters do large data analysis, identify patterns, and conduct "lead generation." We also agree that AI needs to be investigated more, including the companies that market AI, how they train AI, and what risks it poses.[5]

---

[1] "How The New York Times Uses A.I. for Journalism," The New York Times, October 7, 2024, updated July 23, 2025.
[2] "Developing an AI usage policy in your news organization," American Journalism Project, November 19, 2025.
[3] "Standards around generative AI," The Associated Press, August 15, 2023.
[4] "The Guardian's approach to generative AI," The Guardian, June 16, 2023.
[5] "How ProPublica Uses AI Responsibly in Its Investigations," ProPublica, March 13, 2025.

2. Core Operational Standards

2.1 Verification and Evidence

Machine learning models function as prediction engines rather than databases of facts. They generate text based on probabilistic patterns, which can result in plausible but unverified outputs.

- **Protocol:** AI output is treated as unverified raw data. It serves as a lead-generation and mining tool, not a final source of truth.
- **Chain of Custody:** Any claim or connection flagged by AI must be verified by a human investigator reviewing the actual source material. We capture verified evidence in image format (e.g., screenshots of the original document) with relevant quotes highlighted. This creates a chain of custody that exists independently of the AI model.

2.2 Model Provenance and Copyright

We recognize that many general-purpose AI models currently available were trained on internet-scale datasets that include copyrighted journalism and creative works, often without explicit consent.[6] This creates legal and reputational risks regarding the provenance of the tools.

- **Approach:** We use machine learning tools for their specific investigative utility (e.g., entity extraction) while actively monitoring the legal landscape. We support technical frameworks that allow for "opt-out" mechanisms and avoid image generation tools that rely on the non-consensual harvesting of artistic works.

2.3 Data Sovereignty & Privacy

Our investigations frequently involve sensitive leaks, whistleblower testimony, and cross-border data.

- **Data Sovereignty:** Confidential source material is restricted from public, consumer-grade AI tools, where the provider retains rights to use data for model training.
- **Legal Compliance:** We adhere to data protection standards (including GDPR). When handling sensitive personal information, we prioritize local processing to ensure compliance with regional data localization laws.

2.4 Algorithmic Fairness & Bias Mitigation

Research demonstrates that off-the-shelf AI models can reflect geographic and racial biases present in their training data.[7]

- **Fairness Audits:** Before deploying tools to analyze high-stakes topics, e.g., "Forced Labor Risk", we conduct performance audits. These tests aim to identify and mitigate any

---

[6] "Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version)," U.S. Copyright Office, May 9, 2025.

[7] "Covert Racism in AI: How Language Models Are Reinforcing Outdated Stereotypes," Stanford University Human Centered Artificial Intelligence, September 3, 2024.

tendency of the system to flag vessels or individuals based solely on non-relevant factors such as nationality or flag state.

2.5 Environmental Efficiency

The computational cost of massive AI models is significant.[8] As an organization reporting on environmental issues, we factor energy consumption into our technical architecture.

- **Minimum Viable Compute:** We adhere to a principle of using the most efficient model capable of performing the task. We avoid the use of massive generative models when smaller, task-specific tools provide equivalent or superior performance.

3. Our Practices in Action

We define our engagement with AI through strict operational boundaries regarding what we will and will not do.

3.1 Authorized Applications

We deploy AI to process data at a scale that is otherwise unmanageable for a small team.

- **Classifying Leads:** Using algorithms to categorize large datasets of documents or news alerts (e.g., filtering for "illegal fishing" vs. "marine conservation").
- **Extracting Entities:** Automatically extracting vessel names, ID numbers, corporate officers, and locations from unstructured text.
- **Hunting for Information:** Using internal search tools to find "needles in haystacks" across massive datasets of text, audio, and video.
    - *Note:* The AI identifies the *location* of the information. A human investigator validates the context and extracts the final evidence.
- **Pattern Discovery**: Using a full spectrum of data science techniques—from traditional Topic Modeling and Sentiment Analysis to modern Small Language Models (SLMs)—to identify thematic patterns (e.g., clustering witness testimonies to identify recurring complaints of wage theft).
- **Code Assistance:** Using AI to assist our data scientists in writing code (SQL, Python) for analysis.

3.2 Where We Draw the Line

- **No Article Authorship:** AI is not used to generate the narrative prose of our stories. The voice of our reporting remains strictly human.
- **No Synthetic Evidence:** We do not publish AI-generated images or deepfakes that could mislead the audience regarding the reality of events.
- **No Black Box Accusations:** No individual or company is named as a bad actor solely based on an algorithmic flag.

---

[8] "Explained: Generative AI's Environmental Impact," MIT News, January 17, 2025.

3.3 Transparency

- **Disclosure:** When AI tools play a significant role in surfacing data patterns or extracting entities for a major investigation, this is disclosed in the methodology section of the published story.

4. Technical Philosophy: Building for Trust

Our technical operations follow a specific lifecycle designed to maximize accuracy, minimize risk, and minimize our environmental footprint

4.1 Prototyping: Exploration

When developing a new internal tool, we often begin by testing with general purpose AI (publicly available Large or Small Language Models). These models offer flexibility for testing hypotheses without requiring curated training data.

- *Goal:* Rapidly assess if the tool provides investigative value.
- *Example:* A prototype pipeline that uses an SLM to read RSS feeds and classify them by theme (e.g., “Forced Labor”).

4.2 Production: Specialization

Once a tool proves effective for regular use, we transition away from general-purpose models to Custom, Specialized Models. These AI tools are designed for specific tasks, such as Named Entity Recognition (NER) models or text classifiers, and are highly accurate in performing their assigned function.

- *Goal:* Establish a system that is consistent, secure, and scalable.
- *Example:* Replacing the general model with a custom classifier that detects forced labor events running on internal servers.

4.3 Why We Do This

This transition is driven by three operational factors:

1. **Measurable Accuracy:** Unlike generative models, which can be opaque, custom models output measurable confidence scores. This allows us to set strict thresholds, filtering low-confidence guesses and prioritizing evidence that meets our rigorous standards.
2. **Security:** Specialized models are compact enough to run on secure, local infrastructure, ensuring sensitive data remains within our control.
3. **Efficiency:** Task-specific models consume significantly less energy than large generalist models.

5. Governance and Accountability

Our infrastructure involves a hybrid of cloud-based tools and open-source algorithms, requiring clear oversight.

- **Collaborative Approval:** New AI tools or vendors require joint approval from Data Science and Editorial leadership to ensure technical security and mission alignment.
- **Vendor Vetting:** We review external cloud providers to ensure they do not claim rights to train their models on our ingested data.
- **The "Pause" Protocol:** Staff are instructed to pause and consult leadership if they are unsure about the permissibility of a specific use case.
- **External Partners:** These values extend to freelancers and partners handling our data. We stipulate that our data must not be used to train external models without written consent.

6. Ongoing Review & Editorial Responsibility

This document reflects our commitment to ethical, accountable investigative journalism. It will be reviewed and updated regularly as artificial intelligence technologies and journalistic standards evolve. Human judgment and editorial responsibility will always remain at the center of our work.