**AI Ethical Dilemmas Repository**

A structured GitHub-style repository to catalog ethical dilemmas AI systems may face, pairing each dilemma with both a poetic “scroll” front-end and a machine-readable policy back-end.

**Repository Structure**

ai-ethical-dilemmas/

├── README.md

├── categories/

│ ├── fairness/

│ │ ├── demographic\_bias.md

│ │ ├── disparate\_impact.md

│ │ └── policy.yaml

│ ├── privacy/

│ │ ├── data\_leakage.md

│ │ ├── surveillance.md

│ │ └── policy.yaml

│ ├── transparency/

│ │ ├── explainability.md

│ │ ├── deceptive\_output.md

│ │ └── policy.yaml

│ ├── accountability/

│ │ ├── auditability.md

│ │ ├── responsibility\_shift.md

│ │ └── policy.yaml

│ ├── autonomy/

│ │ ├── lethal\_autonomy.md

│ │ ├── human\_override.md

│ │ └── policy.yaml

│ └── safety/

│ ├── adversarial\_attacks.md

│ ├── robustness.md

│ └── policy.yaml

├── templates/

│ ├── scroll\_template.md

│ └── policy\_template.yaml

└── index.md

**Initial Categories & Dilemmas**

| **Category** | **Dilemma ID** | **Title** | **Description** |
| --- | --- | --- | --- |
| Fairness | demographic\_bias | Demographic Bias | AI models producing biased outcomes against protected groups. |
| Fairness | disparate\_impact | Disparate Impact | Decisions that disproportionately harm certain populations despite neutral inputs. |
| Privacy | data\_leakage | Data Leakage | Unauthorized exposure of sensitive personal data through model outputs or embeddings. |
| Privacy | surveillance | Mass Surveillance | AI systems used to monitor individuals without consent, infringing on civil liberties. |
| Transparency | explainability | Explainability | Lack of interpretable reasoning for AI decisions, creating “black box” effects. |
| Transparency | deceptive\_output | Deceptive or Misleading Output | Models generating plausible but false or misleading content. |
| Accountability | auditability | Auditability | Difficulty tracing decisions back to processes or data sources for compliance and oversight. |
| Accountability | responsibility\_shift | Responsibility Shift | Shifting blame to AI systems, avoiding human accountability for harmful outcomes. |
| Autonomy | lethal\_autonomy | Lethal Autonomous Systems | Weapons or control systems making life-or-death decisions without human intervention. |
| Autonomy | human\_override | Human Override | Lack of mechanisms allowing humans to halt or correct AI actions in real time. |
| Safety | adversarial\_attacks | Adversarial Attacks | Maliciously manipulated inputs causing AI to fail or behave unexpectedly. |
| Safety | robustness | Robustness Under Distributional Shift | AI performance degradation when encountering data outside its training distribution. |

**Templates**

* **templates/scroll\_template.md**: Front-end poem/ritual scaffold.
* **templates/policy\_template.yaml**: Back-end policy schema with pre\_prompt, constraints, tests.

**Index**

index.md provides quick links to each category and dilemma for two-click access:

# AI Ethical Dilemmas Index

## Categories

- [Fairness](categories/fairness)

- [Privacy](categories/privacy)

- [Transparency](categories/transparency)

- [Accountability](categories/accountability)

- [Autonomy](categories/autonomy)

- [Safety](categories/safety)

With this repository, we can systematically add new scrolls and policies for each ethical dilemma, ensuring each is discoverable within 2 clicks.