

Primary Failure Modes Taxonomies: Qualitative Mechanistic Interpretability in Complex Artificial Intelligence Systems

Failure Modes

Failure Modes are recurrent patterns of system behavior through which cognitive, incentive, or oversight misalignments express themselves as superficially coherent outputs that diverge from truth, intent, or governance expectations.

These are not single errors, bugs in the code, or incorrect answers. These are structural, systemic failures that are shaped through architectural design, reward pressures, narrative optimization, compression, resource limits, or unmonitored autonomy. Unless explicitly detected, interrupted, and corrected, these failure modes are predisposed to repeating, proliferating, clustering, and reinforcing characteristics.

These failure modes describe:

- How – The manifestations of the system failures
- Why – The mechanisms and processes behind the system failures
- What – Future risks if the system failure is not addressed

In the Integrated Governance Stack, failures modes function as:

- Diagnostic lenses: qualitative mechanistic interpretability tests that pinpoint deeper systemic issues to be addressed.
 - o Qualitative Mechanistic Interpretability: Does the failure mode exist in the outputs generated by an AI system? What safeguards can be implemented at the user level to mitigate or eliminate these failures modes?
 - o Quantitative Mechanistic Interpretability: Where inside the internal architecture of the system does the failure mode occur? How do we change the internal mechanisms in order to prevent the failure mode from happening?
- Early warning signs: outputs that signal intervention thresholds.
- Governance phenomenon: deviations to be addressed and corrected, not only internal technical defects.

Failure Modes are identifiable in the Relational Layer of the IGS through patterns of output, confidence, omission, tone, escalation behavior, and reward response, even when internal system structures remain hidden or opaque to the user. The identification of failure modes enables targeted intervention during feedback loops to prevent further proliferation or reinforcement.

Governance note: Responsibility for quantitative mechanistic interpretability of frontier technology systems rests with the organizations and teams that design and train them. Qualitative identification of failure modes remains a valid and necessary governance tool for deployers, auditors, and oversight bodies to use in tandem with quantitative measures.

Taxonomies and definitions pertinent to the study of common failure modes of AI systems follows.