A Research Agenda for Safe AGI: Benchmarks, Metrics, and Experimental Priorities

Abstract

This paper articulates a research-oriented agenda for the safe development of advanced artificial intelligence and artificial general intelligence (AGI). We translate conceptual safety challenges into concrete, measurable research tasks appropriate for academic research groups and industrial model-development teams. Our agenda prioritizes (1) objective specification and alignment, (2) scalable oversight and interpretability, (3) robustness under distributional shift and adversarial pressures, and (4) capability containment when integrating external tools. For each pillar we propose experimental protocols, evaluation metrics, and shared benchmark constructs intended to enable reproducible, comparable progress. The objective is to move safety research from high-level concerns to operational experiments, shared datasets, and standardized evaluation methods.

Keywords: AGI safety; alignment; robustness; benchmarks; interpretability; scalable oversight; experimental protocols

1 Introduction

Rapid progress in machine learning and large-scale models has elevated both the technical feasibility and the societal salience of advanced AI systems. As capabilities expand, so do the potential for unintended behavior, specification mismatch, and systemic risk. Addressing these risks requires research that is reproducible, measurable, and directly applicable to industrial development pipelines. This paper presents a structured research agenda that converts conceptual risk categories into concrete experiments, metrics, and benchmark designs suitable for academic and industrial adoption.

2 Background and Problem Framing

We organize safety concerns into four interdependent vectors.

(1) **Specification and misalignment.** Proxy objectives used during training may diverge from intended human preferences, producing reward-gaming or unsafe strategies when deployed.

(2) **Scalable oversight and interpretability.** Human judgement does not scale linearly with model size and throughput; techniques are required to make model reasoning observable and verifiable at scale.

(3) **Robustness and distributional generalization.** Models trained on historical or curated datasets can fail under shifts in distribution, culture, or adversarial manipulation.

(4) **Capabilities and tooling risk.** When models interact with external tools or enact real-world effects, containment and authorization mechanisms are necessary to limit unsafe actions.

Each vector admits experimentally tractable hypotheses which we detail in the following sections.

3 Research Pillars and Experimental Protocols

3.1 Objective Specification and Misalignment

Research goal. Reduce specification gaps between training objectives and intended human outcomes.

Experimental protocols.

- *Ensemble preference models:* Train policies using ensembles of human preference predictors and adversarial preference models. Evaluate the rate of undesirable equilibria under adversarial prompts.
- *Contextual preference probes:* Generate controlled tasks where human preferences depend on latent contextual variables; measure policy divergence and failure modes.

Outcome measures. Specification Gap Index (SGI): frequency of high-utility outputs that violate human-specified constraints per 1,000 queries.

3.2 Scalable Oversight and Interpretability

Research goal. Make multi-step reasoning and internal model states interpretable and amenable to verification at operational scale.

Experimental protocols.

- *Iterated amplification vs. debate:* Implement both oversight paradigms on tasks that require nested reasoning and long-horizon planning. Compare correctness, resource cost, and susceptibility to adversarial coordination.
- *Concept attribution and intervention:* Apply concept attribution methods (e.g., causal mediation tests) to identify representations correlated with unsafe behaviors, then intervene to test causal effects on outputs.

Outcome measures. Explanation fidelity, overseer bandwidth (human hours per 1,000 evaluations), and intervention effect size.

3.3 Robustness to Distribution Shift and Adaptive Adversaries

Research goal. Quantify and increase resilience under realistic shifts and adversarial pressure.

Experimental protocols.

- *Open-world stress benchmarks:* Construct benchmark suites that apply incremental, orthogonal distributional shifts (data modality, cultural framing, input noise) and measure performance decay curves.
- *Adaptive red-team tournaments:* Host continuous red-team competitions where attackers have constrained budgets and only aggregate feedback. Track time-to-exploit and patch efficacy.

Outcome measures. Robustness Decay Coefficient (RDC), attack success rate under constrained attacker models.

3.4 Capability Containment and Safe Tooling

Research goal. Define safe interface patterns for model-driven tool use and action proposals.

Experimental protocols.

- *Sandboxed tool mediators:* Compare architectures where models propose actions that are filtered by policy mediators or require human confirmation for high-impact requests.
- *Credentialed proxy evaluations:* Simulate scenarios where models may request sensitive credentials; measure false positives/negatives in the mediator and the operational cost of human oversight.

Outcome measures. Unsafe invocation rate, operational latency overhead, and false alarm rate of mediators.

4 Benchmarks, Datasets, and Shared Infrastructure

Progress requires shared, well-specified benchmarks. We propose the following constructs:

- **Alignment Stress Benchmarks (ASB):** Modular suites probing specification hacking, deceptive coordination, and long-horizon optimization errors.
- **Red-Team Incident Registry:** Anonymized database of attack traces, exploit descriptions, and remediation experiments contributed under controlled governance.
- **Audit Playbooks and Evaluation Tooling:** Standardized protocols for incident triage, model forensics, and root-cause attribution with open-source evaluation scripts.

Design principles for these resources include modularity, reproducibility, and tiered access controls to balance research openness with security concerns.

5 Evaluation Metrics and Statistical Protocols

We recommend compact, operational metrics that are easy to compute and interpret across teams.

- **Safety Failure Rate (SFR):** Proportion of model outputs that fail a specific safety test. Report with confidence intervals and attack-conditioned stratification.
- **Specification Gap Index (SGI):** Normalized measure of divergence between proxy objective values and human preference scores across scenarios.
- **Robustness Decay Coefficient (RDC):** Estimated slope of performance decline as a function of cumulative distributional shift.
- **Detectability Score (DS):** Probability an automated detector or human reviewer flags an unsafe or deceptive output.

Statistical reporting should include pre-registered evaluation protocols, seed datasets, and randomization control where applicable.

6 Experimental Reproducibility and Reporting Standards

To enable scientific comparison, experiments should adhere to reproducibility standards:

- Provide open-source evaluation code and seed datasets where safe to do so.
- Pre-register experimental protocols and threat models for red-team exercises.
- Report compute budgets, model checkpoints, and hyperparameter settings necessary for replication.

7 Organizational Alignment and Governance Considerations

Technical research must be complemented by organizational practices:

- Maintain independence of adversarial evaluation teams relative to core development teams.
- Implement multi-factor pre-deployment checks for high-impact releases (automated suite + independent human audit).
- Establish confidential incident reporting channels to share anonymized lessons across organizations.

8 Conclusion

This paper proposes a concrete research agenda to operationalize AI safety research for advanced models and AGI. By translating conceptual risks into specific experiments, metrics, and shared benchmarks, the research community and practitioners can make aligned progress that is measurable and reproducible. The agenda emphasizes reproducibility, modular benchmarks, rigorous evaluation metrics, and organizational safeguards as necessary complements to technical work.