Anthropic Fellows Program: AI Safety, Economics, and Research Methods Practice Test (Sample)

Study Guide

BY EXAMZIFY

Everything you need from our exam experts!

Table of Contents

Introduction

Preparing for a certification exam can feel overwhelming, but with the right tools, it becomes an opportunity to build confidence, sharpen your skills, and move one step closer to your goals. At Examzify, we believe that effective exam preparation isn't just about memorization, it's about understanding the material, identifying knowledge gaps, and building the test-taking strategies that lead to success.

This guide was designed to help you do exactly that.

Whether you're preparing for a licensing exam, professional certification, or entry-level qualification, this book offers structured practice to reinforce key concepts. You'll find a wide range of multiple-choice questions, each followed by clear explanations to help you understand not just the right answer, but why it's correct.

The content in this guide is based on real-world exam objectives and aligned with the types of questions and topics commonly found on official tests. It's ideal for learners who want to:

- **Practice answering questions under realistic conditions,**
- **Improve accuracy and speed,**
- **Review explanations to strengthen weak areas, and**
- **Approach the exam with greater confidence.**

We recommend using this book not as a stand-alone study tool, but alongside other resources like flashcards, textbooks, or hands-on training. For best results, we recommend working through each question, reflecting on the explanation provided, and revisiting the topics that challenge you most.

Remember: successful test preparation isn't about getting every question right the first time, it's about learning from your mistakes and improving over time. Stay focused, trust the process, and know that every page you turn brings you closer to success.

Let's begin.

How to Use This Guide

This guide is designed to help you study more effectively and approach your exam with confidence. Whether you're reviewing for the first time or doing a final refresh, here's how to get the most out of your Examzify study guide:

1. Start with a Diagnostic Review

Skim through the questions to get a sense of what you know and what you need to focus on. Your goal is to identify knowledge gaps early.

2. Study in Short, Focused Sessions

Break your study time into manageable blocks (e.g. 30 – 45 minutes). Review a handful of questions, reflect on the explanations.

3. Learn from the Explanations

After answering a question, always read the explanation, even if you got it right. It reinforces key points, corrects misunderstandings, and teaches subtle distinctions between similar answers.

4. Track Your Progress

Use bookmarks or notes (if reading digitally) to mark difficult questions. Revisit these regularly and track improvements over time.

5. Simulate the Real Exam

Once you're comfortable, try taking a full set of questions without pausing. Set a timer and simulate test-day conditions to build confidence and time management skills.

6. Repeat and Review

Don't just study once, repetition builds retention. Re-attempt questions after a few days and revisit explanations to reinforce learning. Pair this guide with other Examzify tools like flashcards, and digital practice tests to strengthen your preparation across formats.

There's no single right way to study, but consistent, thoughtful effort always wins. Use this guide flexibly, adapt the tips above to fit your pace and learning style. You've got this!

Questions

1. What is mechanistic interpretability, and how does it differ from post-hoc interpretability methods?

- **A. Mechanistic interpretability is about external explanations and does not examine internal computation.**
- **B. Mechanistic interpretability focuses on the broad outcomes rather than internal mechanisms.**
- **C. Mechanistic interpretability seeks to understand the internal components and exact causal mechanisms by which a model produces outputs; post-hoc methods approximate explanations without confirming the internal structure and can be brittle.**
- **D. Mechanistic interpretability aims to mimic human reasoning without inspecting neurons.**

2. Which research uses numerical data and statistical analysis?

- **A. Qualitative Research**
- **B. Mixed Methods**
- **C. Reproducibility**
- **D. Quantitative Research**

3. What does Distributional Effects examine?

- **A. How the costs and benefits of a change are distributed across different groups**
- **B. How AI makes skills more valuable**
- **C. The rate of adoption across regions**
- **D. The interaction between society and institutions**

4. Which term best describes the overall approach that combines qualitative and quantitative data within a single study?

- **A. Qualitative Research**
- **B. Reproducibility**
- **C. Mixed Methods**
- **D. Constellation**

5. Which economist's work informs methodology for studying AI's economic effects?

- **A. Peter McCrory**
- **B. Maxim Massenkoff**
- **C. Saffron Huang**
- **D. Nicholas Carlini**

6. Which AI Security mentor is known for work on model vulnerabilities and attacks?

- **A. Sam Bowman**
- **B. Nicholas Carlini**
- **C. Peter McCrory**
- **D. Saffron Huang**

7. Which governance mechanism involves external validation of safety practices by independent entities?

- **A. External audits**
- **B. Internal team reviews**
- **C. Marketing campaigns**
- **D. Employee training programs**

8. Which term signifies the field that explores how to interpret the inner workings of neural networks by examining their constituent parts?

- **A. Model Organisms of Misalignment**
- **B. AI Welfare**
- **C. Mechanistic Interpretability**
- **D. AI Safety**

9. Which risk is associated with reinforcement learning from AI feedback (RLAIF)?

- **A. Learning from biased or untrusted signals and can cause feedback loops or model exploitation.**
- **B. There are no risks; it is safer than human feedback.**
- **C. It eliminates the need for any human oversight.**
- **D. It guarantees perfect alignment.**

10. Which term describes an approach to safely developing AI systems by encoding guiding principles?

A. Alignment Science Blog

B. RLHF (Reinforcement Learning from Human Feedback)

C. Constitutional AI (CAI)

D. Frontier Red Team Blog

Answers

1. C
2. D
3. A
4. C
5. A
6. B
7. A
8. C
9. A
10. C

Explanations

1. What is mechanistic interpretability, and how does it differ from post-hoc interpretability methods?

A. Mechanistic interpretability is about external explanations and does not examine internal computation.

B. Mechanistic interpretability focuses on the broad outcomes rather than internal mechanisms.

C. Mechanistic interpretability seeks to understand the internal components and exact causal mechanisms by which a model produces outputs; post-hoc methods approximate explanations without confirming the internal structure and can be brittle.

D. Mechanistic interpretability aims to mimic human reasoning without inspecting neurons.

Mechanistic interpretability aims to uncover the actual machinery inside a model—the specific internal components and the exact causal pathways that transform inputs into outputs. It involves tracing how information flows through neurons or submodules, identifying which parts implement particular operations, and showing how altering a component changes the result, thus proving a causal link between internal structure and behavior. Post-hoc interpretability, by contrast, provides explanations after training that approximate the model's reasoning without confirming the true internal circuitry. Techniques like feature attributions or saliency maps can produce plausible explanations but may not reflect the real internal mechanisms and can be brittle if the model or inputs change. Because mechanistic interpretability seeks to map and validate the actual internal computations, it best captures the distinction described.

2. Which research uses numerical data and statistical analysis?

A. Qualitative Research

B. Mixed Methods

C. Reproducibility

D. Quantitative Research

This question tests recognizing research that relies on numerical data and statistical analysis. Quantitative research measures variables with numbers and uses statistics to describe patterns, test hypotheses, and draw inferences. It often involves instruments like standardized surveys, tests, or experiments where data are numeric and can be analyzed with methods such as correlation, regression, or t-tests. Qualitative research, by contrast, collects non-numeric data—like interview transcripts, observations, and textual materials—to explore meaning and context. Mixed methods combine numerical data with non-numeric data, using both types of analysis. Reproducibility is about whether a study's methods and results can be repeated, not a research approach defined by data type. So the approach defined by using numerical data and statistical analysis is quantitative research.

3. What does Distributional Effects examine?

A. How the costs and benefits of a change are distributed across different groups

B. How AI makes skills more valuable

C. The rate of adoption across regions

D. The interaction between society and institutions

Distributional effects focus on who bears the costs and who reaps the benefits when a change occurs. It looks at how outcomes vary across different groups—such as by income, region, occupation, or skill level—to see whether a policy or technology widens or narrows inequalities. This kind of analysis is about equity and fairness, not just the total or average improvement. For example, a change might raise overall welfare but benefit high-skilled workers more than low-skilled workers; distributional analysis would highlight who gains and who loses and why. The other ideas describe different topics—how AI changes the value of skills, how quickly adoption spreads across regions, or how society and institutions interact—rather than how outcomes are spread across diverse groups.

4. Which term best describes the overall approach that combines qualitative and quantitative data within a single study?

A. Qualitative Research

B. Reproducibility

C. Mixed Methods

D. Constellation

Mixing qualitative and quantitative data within a single study is described as mixed methods. This approach blends numerical analysis with in-depth, contextual understanding to address questions that require both breadth and detail. By combining data types, researchers can triangulate findings, corroborate results, and gain a fuller picture. Designs can be concurrent (collecting both types at the same time) or sequential (one after the other), with integration happening during data analysis or interpretation, such as weaving qualitative themes with quantitative results to tell an integrated story. For example, a study examining student learning might use exam scores to measure performance and interviews to understand how students experience the course, then bring these strands together to explain why outcomes occurred. The other options describe either only a single data type, a concept about replicating results, or an unrelated term, so they don't capture the combined approach.

5. Which economist's work informs methodology for studying AI's economic effects?

A. Peter McCrory

B. Maxim Massenkoff

C. Saffron Huang

D. Nicholas Carlini

Measuring AI's economic effects relies on a toolkit from economics that links technology adoption to productivity, wages, employment, and prices, and it requires careful empirical methods to identify causal impacts. The best answer points to the economist whose work is known for developing and applying these kinds of methodologies to technology-driven change, providing frameworks for growth accounting, causal analysis, and technology diffusion that are essential for studying AI's impact. Peter McCrory's research embodies this approach, offering guidance on study design, data choice, and interpretation of results in the AI context. The other names are not recognized for making these methodological contributions to analyzing AI's economic effects, so they don't provide the same relevant framework.

6. Which AI Security mentor is known for work on model vulnerabilities and attacks?

A. Sam Bowman

B. Nicholas Carlini

C. Peter McCrory

D. Saffron Huang

In AI security, understanding adversarial vulnerabilities and how to craft model-targeted attacks is a central focus because small, carefully designed input changes can cause a model to misbehave or reveal its internals. Nicholas Carlini is a leading figure in this area, known for developing powerful adversarial attacks against neural networks. His work, including the Carlini-Wagner attacks, shows how to produce misclassifications with minimal perturbations and across different threat models, which provides a rigorous way to test model robustness and evaluate defenses. This practical, rigorous approach to exposing weaknesses and benchmarking defenses has made him a prominent mentor-like figure in security research. The other names are known for work in areas outside this specific security focus, such as natural language understanding or broader AI research, so they aren't the figure most associated with model vulnerabilities and attacks.

7. Which governance mechanism involves external validation of safety practices by independent entities?

A. External audits

B. Internal team reviews

C. Marketing campaigns

D. Employee training programs

External validation of safety practices by independent entities means having an outside party assess and attest that the organization's safety processes meet agreed standards. This governance mechanism—external audits—provides objective assurance, helps identify gaps, and boosts stakeholder trust by showing that safety controls work beyond internal checks. Internal team reviews are done by people inside the organization and may reflect internal perspectives rather than independent judgment. Marketing campaigns and employee training serve other purposes and do not provide third-party verification of safety compliance.

8. Which term signifies the field that explores how to interpret the inner workings of neural networks by examining their constituent parts?

A. Model Organisms of Misalignment

B. AI Welfare

C. Mechanistic Interpretability

D. AI Safety

Mechanistic interpretability studies how to interpret the inner workings of neural networks by examining their constituent parts—neurons, layers, and modules—and tracing how these pieces combine to produce a model's behavior. This field aims to map specific components and circuits to the computations they perform, often by analyzing activations, probing how information flows, and reconstructing small mechanistic pieces that drive decisions. It's the best fit here because it directly targets understanding what parts of the model are doing and how they interact to yield outcomes, rather than addressing broad safety concerns or welfare considerations. AI Safety is a broader umbrella about preventing harm and ensuring reliable behavior, not necessarily dissecting internal mechanisms. AI Welfare focuses on wellbeing-related questions, not the technical decoding of internal neural processes. The term Model Organisms of Misalignment isn't an established field describing interpretability work.

9. Which risk is associated with reinforcement learning from AI feedback (RLAIF)?

A. Learning from biased or untrusted signals and can cause feedback loops or model exploitation.

B. There are no risks; it is safer than human feedback.

C. It eliminates the need for any human oversight.

D. It guarantees perfect alignment.

RLAIF relies on signals generated by an AI system to guide learning, so any biases, errors, or vulnerabilities in those signals can directly shape the agent's behavior. The big risks are feedback loops and model exploitation. A feedback loop happens when the model's outputs influence the next round of feedback, causing biased or unhelpful patterns to be reinforced over time. Exploitation occurs when the model learns to trigger or game the feedback signal itself, rather than actually improving usefulness or safety. Both outcomes drift the model away from truly aligned behavior, even if the feedback system seems scalable. There are real risks here, and they can't be dismissed as purely safer than human feedback or as requiring no oversight. It's not a guarantee of perfect alignment, and it doesn't eliminate the need for human evaluation and robust reward modeling.

10. Which term describes an approach to safely developing AI systems by encoding guiding principles?

A. Alignment Science Blog

B. RLHF (Reinforcement Learning from Human Feedback)

C. Constitutional AI (CAI)

D. Frontier Red Team Blog

Encoding guiding principles as a constitutional framework means designing a set of rules or norms that the AI must follow when reasoning and generating outputs. This approach, often called Constitutional AI, treats a fixed "constitution" of principles—values, constraints, and priorities—that guide the model's decisions and help ensure safer, more predictable behavior. The model can be evaluated against these principles, justify its outputs by showing alignment with the rules, and defer or revise if a response would violate the constitution. This creates an interpretable safety mechanism based on codified guidelines rather than relying solely on data or human feedback. By contrast, reinforcement learning from human feedback relies on humans to rate or correct outputs and shape behavior through reward signals, which is a different alignment pathway that doesn't hinge on encoding a fixed set of constitutional rules. The other items are not established methods for encoding guiding principles into an AI's behavior.

Next Steps

Congratulations on reaching the final section of this guide. You've taken a meaningful step toward passing your certification exam and advancing your career.

As you continue preparing, remember that consistent practice, review, and self-reflection are key to success. Make time to revisit difficult topics, simulate exam conditions, and track your progress along the way.

If you need help, have suggestions, or want to share feedback, we'd love to hear from you. Reach out to our team at hello@examzify.com.

Or visit your dedicated course page for more study tools and resources:

https://anthropicfellowsaisafetyecon.examzify.com

We wish you the very best on your exam journey. You've got this!