Byte

https://www.byte.eco/job/23662/

Enforcement Lead Trust & Safety

Description

About Anthropic Anthropic is an AI safety and research company thatâ��s working to build reliable, interpretable, and steerable AI systems. We want AI to be safe and beneficial for our customers and for society as a whole. Our interdisciplinary team has experience across ML, physics, policy, business and product. About the role: As an Enforcement Lead on the Trust and Safety team, you will develop enforcement and monitoring strategies for upholding our Usage Policies. This role sits on the Policy Enforcement team, and will work closely with the Product Policy team to develop deep expertise in our Usage Policy and internal policy standards so that they can be consistently enforced across our suite of products, customers and users. You will be a key driver in shaping and enforcing our policies, making critical decisions that directly impact the safety and integrity of our platform. This role will provide leadership to a cross-functional team conducting risk assessments on new products and features, recommending mitigation strategies, and working closely with Product teams on implementation. IMPORTANT CONTEXT ON THIS ROLE: In this position you may be exposed to and engage with explicit content spanning a range of topics, including those of a sexual, violent, or psychologically disturbing nature. There is also an on-call responsibility across the Policy and Enforcement teams. Responsibilities Drive the development, implementation, and enforcement of AI policies across Anthropicâ��s products, ensuring consistency and effectiveness Build out processes and intake methods to evaluate new products for trust and safety issues, working closely with product teams to establish a robust risk assessment framework Utilize risk analyses and metrics to develop new enforcement mitigation strategies, working closely with Product and Engineering Analyze harmful behavior and identify trends to better understand the threat landscape and help improve our detection methods Make critical policy enforcement decisions, balancing the need for safety with the potential impact on our users and customers Collaborate closely with internal stakeholders, driving alignment on abuse mitigation strategies, and providing strategic guidance to ensure the safe and responsible deployment of AI capabilities Understand and oversee the customer journey across all T&S enforcement actions from detection to user communication to appeals You may be a good fit if you: Understand the challenges and opportunities of operationalizing product policies, including in the content moderation space, and can incorporate this into our enforcement strategy Have experience developing and implementing processes that scale globally in a fast paced environment Think creatively about how to use technology in a way that is safe and beneficial, and ultimately furthers the goal of advancing safe AI systems Foster strong relationships with internal teams, including product, engineering, legal, sales, support, security, and marketing to ensure seamless collaboration and alignment on new product launches. Have a strong ability to navigate ambiguity, build consensus, and execute Are proactive, enjoy identifying policy and process gaps, and owning end to end solutions Have proficiency in SQL and Python Deadline to apply: None. Applications will be reviewed on a rolling basis. Please mention the word **GOOOD** and tag RMjE3LjYxLjIzLjE2MQ== when applying to show you read the job post completely (#RMjE3LjYxLjIzLjE2MQ==). This is a beta feature to avoid spam applicants. Companies can search these words to find applicants that read this and see they’re human.

Contacts

Job listing via RemoteOK.com

Hiring organization
Anthropic

Job Location
Remote

Base Salary
$ 60000 - $ 92500

Date posted
June 2, 2024

Apply Now