Creating a Blueprint for GPT-6

A targeted approach to governance and technical implementation

Jacqueline LESSOFF, M. Besher MASSRI, and Sharon HO

This report was written for the **GPT-6** case of the AI governance hackathon.

Introduction

The recent upsurge in generative AI models, including those developed by OpenAI such as ChatGPT and GPT-5, has garnered heightened attention and discussion of their potential benefits, challenges, and risks. The integration of deep learning has facilitated the development of increasingly intricate models, paving the way for increasingly human-like language processing capabilities. Greater complexity entails more pronounced benefits, challenges, and risks. In order to maximize the benefits, minimize the risks, and ensure the continued advancement of responsible generative AI models, an international, multi-stakeholder approach is required. This report presents a framework for GPT-6 or subsequent iterations of GPT that centers around this collaborative approach and builds on prior technical and policy breakthroughs. Our approach includes the establishment of an independent board that reflects the multi-stakeholder AI landscape, the implementation of a comprehensive risk management framework, and the creation of specific technical approaches such as bias bounties.

While focusing on GPT-6's core development areas, we emphasize aligning AI with human values. As AI systems advance, aligning them with human interests becomes crucial. To accomplish this, we investigate innovative human-AI cooperation throughout our development, testing, and deployment processes to foster collaborative alignment. These techniques are intended to strengthen alignment between AI and human values, improving future AI generations. Our holistic approach to governance, technical execution, risk management, ensures GPT-6's ethical and responsible development to contribute to a more reliable AI ecosystem.

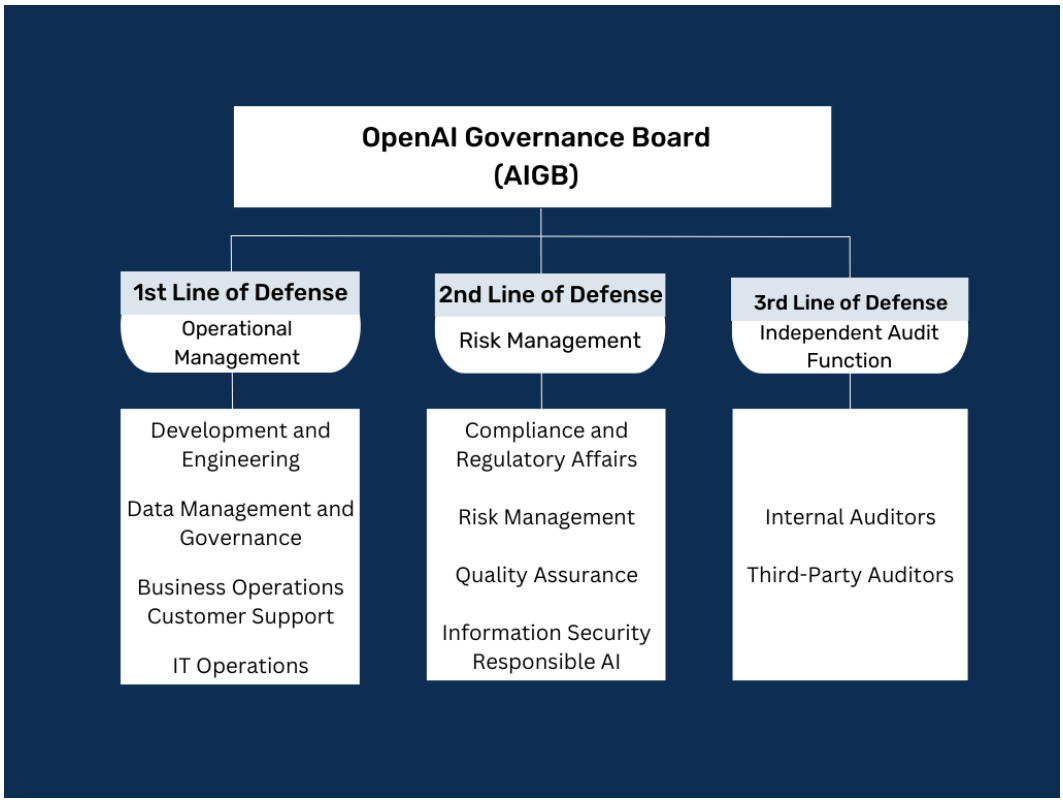
In a pioneering move, OpenAI is revolutionizing the risk management landscape by adopting best practices from the financial services industry, inspired by the robust supervisory approaches employed by central banks. This innovative methodology, never before implemented in the AI sector, is poised to significantly enhance the effectiveness and stability of our organization. Furthermore, our technical approach to stakeholder engagement ensures that all relevant parties – including investors, employees, customers, partners, and

AI Governance Alignment Jam, March 2023

regulators – are actively incorporated into every stage of decision-making, fostering a collaborative ecosystem that drives transparency, accountability, and trust. Our fusion of financial risk management expertise, technical solutions, and comprehensive stakeholder involvement sets a new standard in the AI industry, ensuring that our company remains at the cutting edge of progress, while continually mitigating potential risks and delivering exceptional value for society.

Embedding accountability and transparency into our culture

Improving the risk management structure



To address and manage risks associated with GPT-6 and increase accountability within the organization, **OpenAI will incorporate an improved risk management strategy that formalizes and strengthens the roles of first, second, and third lines of defense**. The first line of defense will include developers, managers, and research scientists who will be responsible for ensuring security in their products. The second line of defense will provide oversight and monitoring of the first line of defense, conducting regular security audits and assessments of the organization's systems and processes to identify

potential vulnerabilities. The compliance team will also conduct regular reviews of the organization's policies and procedures to ensure compliance with relevant regulations and standards. The third line of defense will consist of an independent audit team providing objective and independent assessments of the effectiveness of Open-AI's products and processes.

To oversee this process, **a newly established AI governance board** will provide guidance and oversight to each line of defense. This board will work to ensure that employees are held accountable for identifying, assessing, and controlling risks in their respective areas of work. It will also ensure that risk management and compliance functions have the resources and support needed to develop and implement effective risk management strategies. Finally, the board will ensure that the organization's risk management and control processes are effectively audited and assessed to identify any gaps or deficiencies in the risk management framework, and that third-party auditors remain objective and thorough. By aligning the first, second, and third lines of defense and providing oversight through the AI governance board, OpenAI can effectively manage risks associated with GPT-6 and ensure its continued success.

This new risk management strategy will improve OpenAI's AI accountability to ensure that risks are identified, assessed, and managed appropriately and there are specific defined responsibilities. This alignment will promote transparency and accountability to stakeholders, such as investors and regulators. The AI governance board will also enforce ethical and responsible practices in the development and deployment of AI, reducing the risk profile of GPT-6 and increasing OpenAI's accountability.

Releasing technical documentation

OpenAI also recognizes the importance of transparency in the development and deployment of its models such as GPT-6. For this reason, **OpenAI has decided to release technical detailed information about the algorithms, techniques, and data used by GPT-6**. This information will enable researchers to evaluate the system for potential biases, security vulnerabilities, or ethical concerns.

To ensure that this release does not raise concerns about the potential for reverse engineering or intellectual property theft, we have taken measures to ensure that our system is robust to vulnerabilities. By updating and enhancing our policies and procedures, we ensure a comprehensive approach to security, with measures that encompass physical, network, and application security. Our goal is to ensure that our system is secure and reliable, and that our customers

and stakeholders can trust us to maintain the confidentiality, integrity, and availability of our technology.

Creating a GPT-6 Roadmap

Months 1-3: Preliminary Actions Prior to GPT-6 Development and Deployment

Conducting a strategic review and impact analysis

Prior to the launch of GPT-6, the Openai policy team will conduct a strategic review of their current products of open AI, including a SWOT analysis to identify strengths, weaknesses, opportunities and threats to the company. After understanding the current position of the company, the policy team will conduct an impact report to assess the potential consequences on stakeholders, society, and the environment before rolling out the new model.

Creating an independent governance board

As GPT-6 shows qualities of Artificial General Intelligence, it is imperative to increase transparency and reduce bias by incorporating the input and feedback of a diverse body of experts. As previously announced, OpenAI will create an independent governance board composed of representatives of various sectors, including academia, industry, government, civil society, and user communities. Cultural, racial, and gender diversity will be enforced to ensure fair representation and development of GPT-6 and future products. After a public call for nominations, 12-15 individuals will be chosen by a committee of OpenAI representatives from each business line. The AI Governance board will have specific responsibilities in each phase of product development and roll out, which will be delineated throughout this document and elaborated in OpenAI policies and procedures.

Months 4-6: GPT-6 Development and Policies

Evaluating existing principles and policies

The first task of the newly established governance board will be to review and refine the existing procedures, policies, and principles of OpenAI. Additionally, the board will revise the OpenAI Charter to ensure that the identified human values and goals are robust to challenges presented by GPT-6.. The committee will liaise with external as well as development, legal, and compliance teams to identify opportunities for improvements.

As part of this process, the committee will consider the ethical implications of future OpenAI models and work to ensure that OpenAI's policies and procedures are 'future-proof', and updated to reflect present and anticipated concerns. The committee will also identify opportunities for improvements in areas such as data privacy and security, liability and accountability, and regulatory frameworks.

OpenAI's incident response team will review and propose revisions as necessary to its existing incident response plan, which will include detailed action-plans as well as clear roles and responsibilities for different roles within the organization. The Board will be responsible for reviewing and approving these revisions.

Developing and improving a risk classification framework

In order to ensure that potential long-term impacts and unexpected events are appropriately considered, the second line of defense will develop a risk classification framework that takes into account the different users and applications of GPT-6. This framework will be inspired by existing risk frameworks, such as the NIST AI Risk Classification Framework and the tiered risk framework outlined in the EU AI Act, and will be tailored to the specific features and capabilities of GPT-6.

Once the framework is developed, the second line will assess the possibility of enabling or disabling certain features based on the level of risk associated with each category. In addition, they will identify and incorporate trigger events into the risk registers and lifecycle stages of GPT-6 to ensure that any unintended impacts of risks are quickly identified and addressed. By proactively considering potential long-term impacts and unexpected events, OpenAI will ensure that GPT-6 remains aligned with societal expectations and continues to advance responsible AI development.

Implementing reward engineering

Our development and engineering teams will implement a bottom-up approach to complement the existing top-down-focused approach for aligning the model with human objectives as outlined in the OpenAI Charter. To achieve this, OpenAI plans to adopt inverse reinforcement learning techniques, which involve incorporating human preferences and values from observed behavior to design and update the objective function for the model. Reward engineering will be integrated to encourage the model to behave in a desired manner.

Incorporating reward engineering into the training process allows for more granular control over AI behavior, enabling the system to better align with human values. With this new approach, we can create a more comprehensive alignment framework, ensuring a robust and adaptable AI system. By incorporating this approach, the model will alter its behavior in specific tasks in order to converge with the exhibited values and preferences of human users.

The first and second line of defense will closely monitor the integration of these techniques, making adjustments as necessary to maximize alignment while maintaining high performance levels, with regular reports to the board. As a result, our approach will consist of a comprehensive and multifaceted solution to ensure a reliable and accountable AI ecosystem..

Increasing investment in improving explainable AI

The second line of Defense will examine ways to invest in explainable AI and provide recommendations to the Board. Specifically, it will focus on increasing research and investment for explainability features. These features will serve to better clarify the model training process and provide the development team and risk functions with greater insight into the inner workings of the model. To enhance the explainability of an AI model, one approach is to integrate concept activation vectors, which can identify the specific parts of the input text that are most relevant to the model's predictions. Additionally, attention-based relevance scores can also contribute to explainability by computing a score for each word or phrase in the input text based on its contribution to the model's output. By combining these techniques, we can gain a better understanding of how the AI model arrives at its decisions, which can help build trust and accountability in the system.

In addition, OpenAI plans to collaborate with industry and academia to leverage their knowledge and expertise. OpenAI intends to establish partnerships with academic institutions to fund research and joint projects with other tech companies to develop innovative tools and technologies. OpenAI will also participate in industry initiatives and standards bodies focused on explainability, demonstrating our commitment to advancing transparency and

accountability in AI. OpenAI will continue and strengthen its relationship with the Partnership on AI, and participate more closely on its explainability working group. It will also contribute to the work of the International Organization for Standardisation (ISO), which has developed several standards related to explainability and interpretability of AI, including ISO/IEC TR 24028 and ISO/IEC 23894.

Months 7-9: Deployment and Testing Using Human and AI Solutions

Conducting scenario planning and stress testing

The quality assurance and testing team within the second line of defense will conduct scenario planning and rigorous stress testing exercises to evaluate the AI system's resilience and identify potential vulnerabilities that could affect its reliability and stability under real-world conditions. The stress tests will include attacks on data tampering, adversarial assaults, and social engineering tactics to scrutinize the model's performance and detect any weaknesses.

To further enhance the stress testing process, the quality assurance and testing team will utilize AI solutions using machine learning to identify and anticipate potential issues in the system's behavior. Predictive models will be implemented to simulate the system's performance under different stress conditions, allowing them to identify potential bottlenecks and areas of weakness. By using these methods, the team can anticipate issues such as slow response times, data loss, or system crashes that could occur under extreme conditions. Once these issues have been identified, the development team can take proactive steps to optimize the system's performance and enhance its resilience and robustness under extreme conditions.

The team will also appraise the AI system's capacity to manage substantial query loads and multiple concurrent users to ensure it can meet real-world deployment expectations. Reports to the board will provide insights into the model's strengths and weaknesses, informing targeted improvements aligned with ethical principles and OpenAI objectives.

Conducting beta testing

A first round of testing is conducted to identify bugs, critical issues, and provide performance evaluations. The initial testers will be composed of AI developers, domain experts, and quality assurance professionals, user experience designers, and a small and diverse group of end-users to gauge real-world performance. This process will be overseen by a support team, who will review and summarize feedback to developers.

To identify bugs and critical issues, the development team will perform unit testing and integration testing to ensure that the individual components of the AI system are functioning as expected and can work together seamlessly. They will also conduct functional testing to verify that the system meets the requirements and specifications outlined in the design phase.

Red-teaming

As with prior GPT releases, red-teaming will be conducted to evaluate the resilience and security of AI systems by simulating real-world attack scenarios. The process will involve a team of ethical hackers, cybersecurity experts, and AI researchers with diverse backgrounds who can simulate adversarial attacks on the system.

The blue team, comprising developers, AI engineers, and system administrators are responsible for defending the system against attacks. They work closely with the red team to analyze and respond to potential threats, implementing measures to prevent and mitigate attacks. Reports will be provided to the AI governance board, who plays a critical role in overseeing the red-teaming process by assessing the test results and providing feedback to the development team.

Months 10-12: Post-Deployment Monitoring and Evaluation

Improving evaluation methods and benchmarks

Creating a framework for evaluating the performance and alignment of GPT-6 is crucial for ensuring its overall utility. OpenAI is committed to researching and developing novel evaluation methods that focus on value-sensitive domains and human value-based indicators. To accomplish this, the second line of defense will review and adapt existing benchmarks and evaluation methods to ensure that they are able to address GPT-6's unique characteristics and objectives. Following this review,, the first line of defense will conduct benchmarking and detailed evaluations of the model. These efforts will further enable us to assess GPT-6's alignment with our principles and identify any potential areas of improvement.

To achieve a comprehensive evaluation of our model's capabilities, we will consider a range of performance benchmarks. In addition to natural language understanding and task handling, we will continue to examine the model's proficiency in multitasking and problem-solving. Our updated alignment and safety benchmarks will examine the model's behavior in complex and dynamic environments, and judge its ability to identify and respond to potential harms. A key element in this process will be human-in-the-loop testing, where

the model's performance is assessed in collaboration with human users.. Additionally, qualitative assessments such as narrative understanding evaluations and creativity tests could provide insights into the model's ability to generate innovative and coherent responses. Overall, a combination of quantitative and qualitative assessments can provide a more comprehensive understanding of our model's capabilities and limitations.

Implementing an independent auditing framework

Independent auditing and continuous monitoring are critical components of ensuring the responsible development and deployment of GPT-6. By engaging independent auditors, OpenAI can gain valuable insights and unbiased evaluations of the system's performance and adherence to ethical guidelines. These auditors can identify potential risks, biases, or unintended consequences associated with GPT-6's deployment and recommend improvements or corrective measures.

OpenAI will implement a framework for internal and third party continuous monitoring to track the real-world impact of GPT-6, ensuring that the system evolves in response to new challenges and remains aligned with societal expectations. It is essential to establish a transparent and structured communication channel between auditors and the second line of defense to ensure that all identified issues are addressed and resolved promptly. This will require the implementation of a scheduled reporting mechanism to facilitate communication among independent auditors and the first and second lines of defense. The governance board will provide oversight and guidance to ensure that the independent auditors are objective, thorough, and aligned with the overall strategy and objectives of the organization.

Establishing bias bounties

The governance board will establish bias bounties to proactively identify and address potential biases and vulnerabilities in GPT-6. Bias bounties incentivize the public to discover and report biases or other ethical concerns in GPT-6's outputs, enabling continuous improvement in the system's fairness and inclusivity. By implementing bias bounties, and encouraging participation from our users, OpenAI further increases transparency and inclusivity in the development and deployment of AI systems. We will be launching an interface for GPT-6 that allows anyone to submit areas where the system may be underperforming and propose improvements. This will enable a broader range of stakeholders to participate in the development and improvement of the system, helping to identify potential biases and improve the overall performance of GPT-6. As a bounty, users who submit helpful reports will be awarded GPT credits.

To further streamline the bias bounty process and identify high-risk bounties, we plan to incorporate AI systems to cluster and categorize reported issues based on their severity, potential impact, and likelihood of occurrence. By analyzing patterns and trends in reported issues, the AI system can identify high-risk bounties that require immediate attention and prioritization. This will enable us to efficiently allocate resources and address the most critical issues first, while also providing transparency and accountability to stakeholders. With this approach, we aim to continuously improve the fairness and inclusivity of GPT-6 while also ensuring that the system remains secure and reliable.

The second line of defense, consisting of risk management and compliance functions, will play a crucial role in identifying and addressing issues raised by bias bounties. Once a bias bounty is flagged, the risk management and compliance functions will conduct a thorough investigation to determine the nature and severity of the issue. They will then categorize the issue based on its potential impact and severity, and communicate this information to the first line of defense, consisting of developers, managers, and research scientists. The first line of defense will work to address the issue and develop a solution, while the second line of defense will monitor progress and ensure that the solution is effective. The risk management and compliance functions will also conduct regular audits and assessments to ensure that the established policies and procedures for addressing bias bounties are being followed effectively.

Months 13+: Model Release and Post-Release Monitoring and Improvement

Releasing the model

The AI model will be released alongside monitoring tools that will track the AI system's performance metrics such as response time, server load, and error rates. Additionally, custom solutions are implemented to monitor AI-specific metrics, such as accuracy, F1 score, BLEU, and Perplexity, to track and evaluate the model's effectiveness in various tasks. Our user experience team will also collect immediate user feedback regarding the AI model's output, usability, and any unexpected issues.

Following the release, OpenAI will continue to focus on monitoring, evaluation, and improvement. Our development teams will review the collected performance metrics, user feedback, and test results to identify areas of improvement. Based on these findings, the teams will adjust the model to improve its performance, reduce biases, and better cater user needs.

Depending on the identified issues and improvement areas, the AI model may undergo fine-tuning to enhance and improve its capabilities. As such

changes are made, the associated documentation will be updated to reflect the latest model architecture, training data, algorithms, and system outputs. In addition, the AI development team and AI governance board will hold regular meetings to discuss these developments.

Implementing bias bounties

We plan to launch our Bias Bounties platform simultaneously with the release of GPT-6. We will be making adjustments such as updating the user interface, deploying new backend services, and revising user documentation to educate users about the feature and its intended purpose. To promote the Bias Bounties feature and encourage participation, we will launch targeted marketing campaigns and announcements to raise awareness and engage users.

Additionally, a dedicated team within the second line of defense will closely monitor user engagement with the feature, tracking metrics such as the number of reported biases, the accuracy of reported biases, and user satisfaction with the reward structure. Based on the results, the second line will communicate with the dedicated AI development in order to suggest improvements to the new feature such as adjusting the reward structure, improving the user interface, or modifying the bias verification mechanism.

Conclusion

Alongside the release of GPT-6, OpenAI will introduce new innovative features to enhance our risk profile and address anticipated challenges.. Our approach involves drawing parallels between finance and AI regulation to tailor existing policy and risk management approaches in finance to AI. This will involve incorporating risk management strategies such as risk identification, assessment, monitoring, and continuous improvement into the development and deployment of GPT-6, as well as the creation of an AI governance board to provide expert guidance and oversight

The roll-out of GPT-6 will be a collaborative effort that involves multiple stakeholders and a phased approach. The successful deployment of GPT-6 will require ongoing collaboration and continuous improvement to ensure that the model meets the highest ethical and transparency standards, and that it contributes to advancing responsible AI practices across the industry.

Source:

https://www.bis.org/fsi/fsipapers11.pdf

https://www.newyorkfed.org/banking/supervisionregulate

https://arxiv.org/abs/2201.10408

https://arxiv.org/pdf/2302.08500.pdf

https://openai.com/blog/our-approach-to-alignment-research

https://csrc.nist.gov/projects/risk-management/about-rmf

https://oecd.ai/en/classification/

https://oecd.ai/en/ai-principles

https://csrc.nist.gov/projects/risk-management/about-rmf

https://arxiv.org/abs/2202.03286

https://openai.com/charter

https://www.danieldewey.net/reward-engineering-principle.pdf