MARK R. WARNER
VIRGINIA

United States Senate
WASHINGTON, DC 20510–4606

COMMITTEES:
FINANCE
BANKING, HOUSING, AND URBAN AFFAIRS
BUDGET
INTELLIGENCE
RULES AND ADMINISTRATION

April 26, 2023

Dr. Dario Amodei
Chief Executive Officer
Anthropic
548 Market St, PMB 90375
San Francisco, CA 94104

Dear Dr. Amodei,

I write today regarding the need to prioritize security in the design and development of artificial intelligence (AI) systems. As companies like yours make rapid advancements in AI, we must acknowledge the security risks inherent in this technology and ensure AI development and adoption proceeds in a responsible and secure way. While public concern about the safety and security of AI has been on the rise, I know that work on AI security is not new. However, with the increasing use of AI across large swaths of our economy, and the possibility for large language models to be steadily integrated into a range of existing systems, from healthcare to finance sectors, I see an urgent need to underscore the importance of putting security at the forefront of your work. Beyond industry commitments, however, it is also clear that some level of regulation is necessary in this field.

I recognize the important work you and your colleagues are doing to advance AI. As a leading company in this emerging technology, I believe you have a responsibility to ensure that your technology products and systems are secure. I have long advocated for incorporating security-by-design, as we have found time and again that failing to consider security early in the product development lifecycle leads to more costly and less effective security. Instead, incorporating security upfront can reduce costs[1] and risks[2]. Moreover, the last five years have demonstrated that the ways in which the speed, scale, and excitement associated with new technologies have frequently obscured the shortcomings of their creators in anticipating the harmful effects of their use. AI capabilities hold enormous potential; however, we must ensure that they do not advance without appropriate safeguards and regulation.

While it is important to apply many of the same security principles we associate with traditional computing services and devices, AI presents a new set of security concerns that are distinct from

[1] Maurice Dawson et al., "Integrating Software Assurance into the Software Development Life Cycle (SDLC)," Journal of Information Systems Technology and Planning (2010), Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC
[2] "U.S. and International Partners Publish Secure-by-Design and -Default Principles and Approaches," Cybersecurity and Infrastructure Security Agency (April 13, 2023), https://www.cisa.gov/news-events/news/us-and-international-partners-publish-secure-design-and-default-principles-and-approaches

traditional software vulnerabilities. Some of the AI-specific security risks that I am concerned about include the origin, quality, and accuracy of input data (data supply chain)[3], tampering with training data (data poisoning attacks)[4], and inputs to models that intentionally cause them to make mistakes (adversarial examples)[5]. Each of these risks further highlighting the need for secure, quality data inputs. Broadly speaking, these techniques can effectively defeat or degrade the integrity, security, or performance of an AI system (including the potential confidentiality of its training data). As leading models are increasingly integrated into larger systems, often without fully mapping dependencies and downstream implications, the effects of adversarial attacks on AI systems are only magnified.

In addition to those risks, I also have concerns regarding bias, trustworthiness, and potential misuse or malicious use of AI systems. In the last six months, we have seen open source researchers repeatedly exploit a number of prominent, publicly-accessible generative models – crafting a range of clever (and often foreseeable) prompts to easily circumvent a system's rules. Examples include using widely-adopted models to generate malware[6], craft increasingly sophisticated phishing techniques[7], contribute to disinformation[8], and provide harmful information[9]. It is imperative that we address threats to not only digital security, but also threats to physical security and political security.[10]

In light of this, I am interested in learning about the measures that your company is taking to ensure the security of its AI systems. I request that you provide answers to the following questions no later than May 26, 2023.

Questions:

1. Can you provide an overview of your company's security approach or strategy?
2. What limits do you enforce on third-party access to your model and how do you actively monitor for non-compliant uses?

---

[3] "OWASP AI Security and Privacy Guide," OWASP Foundation, https://owasp.org/www-project-ai-security-and-privacy-guide/

[4] Fahri Anıl Yerlikaya, Şerif Bahtiyar, "Data poisoning attacks against machine learning algorithms", Expert Systems with Applications, Volume 208, (July 18, 2022). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417422012933

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial Examples in the Physical World," Google, Inc. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf

[6] Dan Goodin, "Hackers Are Selling A Service that Bypasses ChatGPT Restrictions on Malware," Ars Technica (February 8, 2023) https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/

[7] Lily Hay Newman, "AI Wrote Better Phishing Emails Than Humans in a Recent Test," Wired (August 7, 2021), https://www.wired.com/story/ai-phishing-emails/

[8] Tiffany Hsu, Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times (February 13, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[9] "GPT-4 Jailbreak and Hacking via RabbitHole attack, Prompt injection, Content moderation bypass and Weaponizing AI," Adversa AI (March 15, 2023), https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/

[10] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (February, 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217

3. Are you participating in third party (internal or external) test & evaluation, verification & validation of your systems?
4. What steps have you taken to ensure that you have secure and accurate data inputs and outputs? Have you provided comprehensive and accurate documentation of your training data to downstream users to allow them to evaluate whether your model is appropriate for their use?
5. Do you provide complete and accurate documentation of your model to commercial users? Which documentation standards or procedures do you rely on?
6. What kind of input sanitization techniques do you implement to ensure that your systems are not susceptible to prompt injection techniques that pose underlying system risks?
7. How are you monitoring and auditing your systems to detect and mitigate security breaches?
8. Can you explain the security measures that you take to prevent unauthorized access to your systems and models?
9. How do you protect your systems against potential breaches or cyberattacks? Do you have a plan in place to respond to a potential security incident? What is your process for alerting users that have integrated your model into downstream systems?
10. What is your process for ensuring the privacy of sensitive or personal information you that your system uses?
11. Can you describe how your company has handled past security incidents?
12. What security standards, if any, are you adhering to? Are you using NIST's AI Risk Management Framework?[11]
13. Is your company participating in the development of technical standards related to AI and AI security?
14. How are you ensuring that your company continues to be knowledgeable about evolving security best practices and risks?
15. How is your company addressing concerns about AI trustworthiness, including potential algorithmic bias and misuse or malicious use of AI?
16. Have you identified any security challenges unique to AI that you believe policymakers should address?

Thank you for your attention to these important matters and I look forward to your response.

Sincerely,

Mark R. Warner
United States Senator

[11] "AI Risk Management Framework," NIST (July 12, 2021), https://www.nist.gov/itl/ai-risk-management-framework

MARK R. WARNER
VIRGINIA

United States Senate
WASHINGTON, DC 20510–4606

COMMITTEES:
FINANCE
BANKING, HOUSING, AND URBAN AFFAIRS
BUDGET
INTELLIGENCE
RULES AND ADMINISTRATION

April 26, 2023

Mr. Tim Cook
Chief Executive Officer
Apple
One Apple Park Way
Cupertino, CA 95014

Dear Mr. Cook,

I write today regarding the need to prioritize security in the design and development of artificial intelligence (AI) systems. As companies like yours make rapid advancements in AI, we must acknowledge the security risks inherent in this technology and ensure AI development and adoption proceeds in a responsible and secure way. While public concern about the safety and security of AI has been on the rise, I know that work on AI security is not new. However, with the increasing use of AI across large swaths of our economy, and the possibility for large language models to be steadily integrated into a range of existing systems, from healthcare to finance sectors, I see an urgent need to underscore the importance of putting security at the forefront of your work. Beyond industry commitments, however, it is also clear that some level of regulation is necessary in this field.

I recognize the important work you and your colleagues are doing to advance AI. As a leading company in this emerging technology, I believe you have a responsibility to ensure that your technology products and systems are secure. I have long advocated for incorporating security-by-design, as we have found time and again that failing to consider security early in the product development lifecycle leads to more costly and less effective security. Instead, incorporating security upfront can reduce costs[1] and risks[2]. Moreover, the last five years have demonstrated that the ways in which the speed, scale, and excitement associated with new technologies have frequently obscured the shortcomings of their creators in anticipating the harmful effects of their use. AI capabilities hold enormous potential; however, we must ensure that they do not advance without appropriate safeguards and regulation.

While it is important to apply many of the same security principles we associate with traditional computing services and devices, AI presents a new set of security concerns that are distinct from

[1] Maurice Dawson et al., "Integrating Software Assurance into the Software Development Life Cycle (SDLC)," Journal of Information Systems Technology and Planning (2010), Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC
[2] "U.S. and International Partners Publish Secure-by-Design and -Default Principles and Approaches," Cybersecurity and Infrastructure Security Agency (April 13, 2023), https://www.cisa.gov/news-events/news/us-and-international-partners-publish-secure-design-and-default-principles-and-approaches

traditional software vulnerabilities. Some of the AI-specific security risks that I am concerned about include the origin, quality, and accuracy of input data (data supply chain)[3], tampering with training data (data poisoning attacks)[4], and inputs to models that intentionally cause them to make mistakes (adversarial examples)[5]. Each of these risks further highlighting the need for secure, quality data inputs. Broadly speaking, these techniques can effectively defeat or degrade the integrity, security, or performance of an AI system (including the potential confidentiality of its training data). As leading models are increasingly integrated into larger systems, often without fully mapping dependencies and downstream implications, the effects of adversarial attacks on AI systems are only magnified.

In addition to those risks, I also have concerns regarding bias, trustworthiness, and potential misuse or malicious use of AI systems. In the last six months, we have seen open source researchers repeatedly exploit a number of prominent, publicly-accessible generative models – crafting a range of clever (and often foreseeable) prompts to easily circumvent a system's rules. Examples include using widely-adopted models to generate malware[6], craft increasingly sophisticated phishing techniques[7], contribute to disinformation[8], and provide harmful information[9]. It is imperative that we address threats to not only digital security, but also threats to physical security and political security.[10]

In light of this, I am interested in learning about the measures that your company is taking to ensure the security of its AI systems. I request that you provide answers to the following questions no later than May 26, 2023.

Questions:

1. Can you provide an overview of your company's security approach or strategy?
2. What limits do you enforce on third-party access to your model and how do you actively monitor for non-compliant uses?

---

[3] "OWASP AI Security and Privacy Guide," OWASP Foundation, https://owasp.org/www-project-ai-security-and-privacy-guide/

[4] Fahri Anıl Yerlikaya, Şerif Bahtiyar, "Data poisoning attacks against machine learning algorithms", Expert Systems with Applications, Volume 208, (July 18, 2022). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417422012933

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial Examples in the Physical World," Google, Inc. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf

[6] Dan Goodin, "Hackers Are Selling A Service that Bypasses ChatGPT Restrictions on Malware," Ars Technica (February 8, 2023) https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/

[7] Lily Hay Newman, "AI Wrote Better Phishing Emails Than Humans in a Recent Test," Wired (August 7, 2021), https://www.wired.com/story/ai-phishing-emails/

[8] Tiffany Hsu, Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times (February 13, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[9] "GPT-4 Jailbreak and Hacking via RabbitHole attack, Prompt injection, Content moderation bypass and Weaponizing AI," Adversa AI (March 15, 2023), https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/

[10] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (February, 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217

3. Are you participating in third party (internal or external) test & evaluation, verification & validation of your systems?
4. What steps have you taken to ensure that you have secure and accurate data inputs and outputs? Have you provided comprehensive and accurate documentation of your training data to downstream users to allow them to evaluate whether your model is appropriate for their use?
5. Do you provide complete and accurate documentation of your model to commercial users? Which documentation standards or procedures do you rely on?
6. What kind of input sanitization techniques do you implement to ensure that your systems are not susceptible to prompt injection techniques that pose underlying system risks?
7. How are you monitoring and auditing your systems to detect and mitigate security breaches?
8. Can you explain the security measures that you take to prevent unauthorized access to your systems and models?
9. How do you protect your systems against potential breaches or cyberattacks? Do you have a plan in place to respond to a potential security incident? What is your process for alerting users that have integrated your model into downstream systems?
10. What is your process for ensuring the privacy of sensitive or personal information you that your system uses?
11. Can you describe how your company has handled past security incidents?
12. What security standards, if any, are you adhering to? Are you using NIST's AI Risk Management Framework?[11]
13. Is your company participating in the development of technical standards related to AI and AI security?
14. How are you ensuring that your company continues to be knowledgeable about evolving security best practices and risks?
15. How is your company addressing concerns about AI trustworthiness, including potential algorithmic bias and misuse or malicious use of AI?
16. Have you identified any security challenges unique to AI that you believe policymakers should address?

Thank you for your attention to these important matters and I look forward to your response.

Sincerely,

Mark R. Warner
United States Senator

[11] "AI Risk Management Framework," NIST (July 12, 2021), https://www.nist.gov/itl/ai-risk-management-framework

MARK R. WARNER
VIRGINIA

United States Senate
WASHINGTON, DC 20510–4606

COMMITTEES:
FINANCE
BANKING, HOUSING, AND URBAN AFFAIRS
BUDGET
INTELLIGENCE
RULES AND ADMINISTRATION

April 26, 2023

Mr. Sundar Pichai
Chief Executive Officer
Google
1600 Amphitheater Parkway
Mountain View, CA 94043

Dear Mr. Pichai,

I write today regarding the need to prioritize security in the design and development of artificial intelligence (AI) systems. As companies like yours make rapid advancements in AI, we must acknowledge the security risks inherent in this technology and ensure AI development and adoption proceeds in a responsible and secure way. While public concern about the safety and security of AI has been on the rise, I know that work on AI security is not new. However, with the increasing use of AI across large swaths of our economy, and the possibility for large language models to be steadily integrated into a range of existing systems, from healthcare to finance sectors, I see an urgent need to underscore the importance of putting security at the forefront of your work. Beyond industry commitments, however, it is also clear that some level of regulation is necessary in this field.

I recognize the important work you and your colleagues are doing to advance AI. As a leading company in this emerging technology, I believe you have a responsibility to ensure that your technology products and systems are secure. I have long advocated for incorporating security-by-design, as we have found time and again that failing to consider security early in the product development lifecycle leads to more costly and less effective security. Instead, incorporating security upfront can reduce costs[1] and risks[2]. Moreover, the last five years have demonstrated that the ways in which the speed, scale, and excitement associated with new technologies have frequently obscured the shortcomings of their creators in anticipating the harmful effects of their use. AI capabilities hold enormous potential; however, we must ensure that they do not advance without appropriate safeguards and regulation.

While it is important to apply many of the same security principles we associate with traditional computing services and devices, AI presents a new set of security concerns that are distinct from

[1] Maurice Dawson et al., "Integrating Software Assurance into the Software Development Life Cycle (SDLC)," Journal of Information Systems Technology and Planning (2010), Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC
[2] "U.S. and International Partners Publish Secure-by-Design and -Default Principles and Approaches," Cybersecurity and Infrastructure Security Agency (April 13, 2023), https://www.cisa.gov/news-events/news/us-and-international-partners-publish-secure-design-and-default-principles-and-approaches

traditional software vulnerabilities. Some of the AI-specific security risks that I am concerned about include the origin, quality, and accuracy of input data (data supply chain)[3], tampering with training data (data poisoning attacks)[4], and inputs to models that intentionally cause them to make mistakes (adversarial examples)[5]. Each of these risks further highlighting the need for secure, quality data inputs. Broadly speaking, these techniques can effectively defeat or degrade the integrity, security, or performance of an AI system (including the potential confidentiality of its training data). As leading models are increasingly integrated into larger systems, often without fully mapping dependencies and downstream implications, the effects of adversarial attacks on AI systems are only magnified.

In addition to those risks, I also have concerns regarding bias, trustworthiness, and potential misuse or malicious use of AI systems. In the last six months, we have seen open source researchers repeatedly exploit a number of prominent, publicly-accessible generative models – crafting a range of clever (and often foreseeable) prompts to easily circumvent a system's rules. Examples include using widely-adopted models to generate malware[6], craft increasingly sophisticated phishing techniques[7], contribute to disinformation[8], and provide harmful information[9]. It is imperative that we address threats to not only digital security, but also threats to physical security and political security.[10]

In light of this, I am interested in learning about the measures that your company is taking to ensure the security of its AI systems. I request that you provide answers to the following questions no later than May 26, 2023.

Questions:

1. Can you provide an overview of your company's security approach or strategy?
2. What limits do you enforce on third-party access to your model and how do you actively monitor for non-compliant uses?

---

[3] "OWASP AI Security and Privacy Guide," OWASP Foundation, https://owasp.org/www-project-ai-security-and-privacy-guide/

[4] Fahri Anıl Yerlikaya, Şerif Bahtiyar, "Data poisoning attacks against machine learning algorithms", Expert Systems with Applications, Volume 208, (July 18, 2022). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417422012933

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial Examples in the Physical World," Google, Inc. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf

[6] Dan Goodin, "Hackers Are Selling A Service that Bypasses ChatGPT Restrictions on Malware," Ars Technica (February 8, 2023) https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/

[7] Lily Hay Newman, "AI Wrote Better Phishing Emails Than Humans in a Recent Test," Wired (August 7, 2021), https://www.wired.com/story/ai-phishing-emails/

[8] Tiffany Hsu, Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times (February 13, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[9] "GPT-4 Jailbreak and Hacking via RabbitHole attack, Prompt injection, Content moderation bypass and Weaponizing AI," Adversa AI (March 15, 2023), https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/

[10] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (February, 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217

3. Are you participating in third party (internal or external) test & evaluation, verification & validation of your systems?
4. What steps have you taken to ensure that you have secure and accurate data inputs and outputs? Have you provided comprehensive and accurate documentation of your training data to downstream users to allow them to evaluate whether your model is appropriate for their use?
5. Do you provide complete and accurate documentation of your model to commercial users? Which documentation standards or procedures do you rely on?
6. What kind of input sanitization techniques do you implement to ensure that your systems are not susceptible to prompt injection techniques that pose underlying system risks?
7. How are you monitoring and auditing your systems to detect and mitigate security breaches?
8. Can you explain the security measures that you take to prevent unauthorized access to your systems and models?
9. How do you protect your systems against potential breaches or cyberattacks? Do you have a plan in place to respond to a potential security incident? What is your process for alerting users that have integrated your model into downstream systems?
10. What is your process for ensuring the privacy of sensitive or personal information you that your system uses?
11. Can you describe how your company has handled past security incidents?
12. What security standards, if any, are you adhering to? Are you using NIST's AI Risk Management Framework?[11]
13. Is your company participating in the development of technical standards related to AI and AI security?
14. How are you ensuring that your company continues to be knowledgeable about evolving security best practices and risks?
15. How is your company addressing concerns about AI trustworthiness, including potential algorithmic bias and misuse or malicious use of AI?
16. Have you identified any security challenges unique to AI that you believe policymakers should address?

Thank you for your attention to these important matters and I look forward to your response.

Sincerely,

Mark R. Warner
United States Senator

[11] "AI Risk Management Framework," NIST (July 12, 2021), https://www.nist.gov/itl/ai-risk-management-framework

MARK R. WARNER
VIRGINIA

United States Senate
WASHINGTON, DC 20510–4606

COMMITTEES:
FINANCE
BANKING, HOUSING, AND URBAN AFFAIRS
BUDGET
INTELLIGENCE
RULES AND ADMINISTRATION

April 26, 2023

Mr. Mark Zuckerberg
Chief Executive Officer
Meta Platforms, Inc.
1 Hacker Way
Menlo Park, CA 94025

Dear Mr. Zuckerberg,

I write today regarding the need to prioritize security in the design and development of artificial intelligence (AI) systems. As companies like yours make rapid advancements in AI, we must acknowledge the security risks inherent in this technology and ensure AI development and adoption proceeds in a responsible and secure way. While public concern about the safety and security of AI has been on the rise, I know that work on AI security is not new. However, with the increasing use of AI across large swaths of our economy, and the possibility for large language models to be steadily integrated into a range of existing systems, from healthcare to finance sectors, I see an urgent need to underscore the importance of putting security at the forefront of your work. Beyond industry commitments, however, it is also clear that some level of regulation is necessary in this field.

I recognize the important work you and your colleagues are doing to advance AI. As a leading company in this emerging technology, I believe you have a responsibility to ensure that your technology products and systems are secure. I have long advocated for incorporating security-by-design, as we have found time and again that failing to consider security early in the product development lifecycle leads to more costly and less effective security. Instead, incorporating security upfront can reduce costs[1] and risks[2]. Moreover, the last five years have demonstrated that the ways in which the speed, scale, and excitement associated with new technologies have frequently obscured the shortcomings of their creators in anticipating the harmful effects of their use. AI capabilities hold enormous potential; however, we must ensure that they do not advance without appropriate safeguards and regulation.

While it is important to apply many of the same security principles we associate with traditional computing services and devices, AI presents a new set of security concerns that are distinct from

[1] Maurice Dawson et al., "Integrating Software Assurance into the Software Development Life Cycle (SDLC)," Journal of Information Systems Technology and Planning (2010), Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC
[2] "U.S. and International Partners Publish Secure-by-Design and -Default Principles and Approaches," Cybersecurity and Infrastructure Security Agency (April 13, 2023), https://www.cisa.gov/news-events/news/us-and-international-partners-publish-secure-design-and-default-principles-and-approaches

traditional software vulnerabilities. Some of the AI-specific security risks that I am concerned about include the origin, quality, and accuracy of input data (data supply chain)[3], tampering with training data (data poisoning attacks)[4], and inputs to models that intentionally cause them to make mistakes (adversarial examples)[5]. Each of these risks further highlighting the need for secure, quality data inputs. Broadly speaking, these techniques can effectively defeat or degrade the integrity, security, or performance of an AI system (including the potential confidentiality of its training data). As leading models are increasingly integrated into larger systems, often without fully mapping dependencies and downstream implications, the effects of adversarial attacks on AI systems are only magnified.

In addition to those risks, I also have concerns regarding bias, trustworthiness, and potential misuse or malicious use of AI systems. In the last six months, we have seen open source researchers repeatedly exploit a number of prominent, publicly-accessible generative models – crafting a range of clever (and often foreseeable) prompts to easily circumvent a system's rules. Examples include using widely-adopted models to generate malware[6], craft increasingly sophisticated phishing techniques[7], contribute to disinformation[8], and provide harmful information[9]. It is imperative that we address threats to not only digital security, but also threats to physical security and political security.[10]

In light of this, I am interested in learning about the measures that your company is taking to ensure the security of its AI systems. I request that you provide answers to the following questions no later than May 26, 2023.

Questions:

1. Can you provide an overview of your company's security approach or strategy?
2. What limits do you enforce on third-party access to your model and how do you actively monitor for non-compliant uses?

---

[3] "OWASP AI Security and Privacy Guide," OWASP Foundation, https://owasp.org/www-project-ai-security-and-privacy-guide/

[4] Fahri Anıl Yerlikaya, Şerif Bahtiyar, "Data poisoning attacks against machine learning algorithms", Expert Systems with Applications, Volume 208, (July 18, 2022). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417422012933

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial Examples in the Physical World," Google, Inc. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf

[6] Dan Goodin, "Hackers Are Selling A Service that Bypasses ChatGPT Restrictions on Malware," Ars Technica (February 8, 2023) https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/

[7] Lily Hay Newman, "AI Wrote Better Phishing Emails Than Humans in a Recent Test," Wired (August 7, 2021), https://www.wired.com/story/ai-phishing-emails/

[8] Tiffany Hsu, Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times (February 13, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[9] "GPT-4 Jailbreak and Hacking via RabbitHole attack, Prompt injection, Content moderation bypass and Weaponizing AI," Adversa AI (March 15, 2023), https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/

[10] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (February, 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217

3. Are you participating in third party (internal or external) test & evaluation, verification & validation of your systems?
4. What steps have you taken to ensure that you have secure and accurate data inputs and outputs? Have you provided comprehensive and accurate documentation of your training data to downstream users to allow them to evaluate whether your model is appropriate for their use?
5. Do you provide complete and accurate documentation of your model to commercial users? Which documentation standards or procedures do you rely on?
6. What kind of input sanitization techniques do you implement to ensure that your systems are not susceptible to prompt injection techniques that pose underlying system risks?
7. How are you monitoring and auditing your systems to detect and mitigate security breaches?
8. Can you explain the security measures that you take to prevent unauthorized access to your systems and models?
9. How do you protect your systems against potential breaches or cyberattacks? Do you have a plan in place to respond to a potential security incident? What is your process for alerting users that have integrated your model into downstream systems?
10. What is your process for ensuring the privacy of sensitive or personal information you that your system uses?
11. Can you describe how your company has handled past security incidents?
12. What security standards, if any, are you adhering to? Are you using NIST's AI Risk Management Framework?[11]
13. Is your company participating in the development of technical standards related to AI and AI security?
14. How are you ensuring that your company continues to be knowledgeable about evolving security best practices and risks?
15. How is your company addressing concerns about AI trustworthiness, including potential algorithmic bias and misuse or malicious use of AI?
16. Have you identified any security challenges unique to AI that you believe policymakers should address?

Thank you for your attention to these important matters and I look forward to your response.

Sincerely,

Mark R. Warner
United States Senator

[11] "AI Risk Management Framework," NIST (July 12, 2021), https://www.nist.gov/itl/ai-risk-management-framework

MARK R. WARNER
VIRGINIA

United States Senate
WASHINGTON, DC 20510–4606

COMMITTEES:
FINANCE
BANKING, HOUSING, AND URBAN AFFAIRS
BUDGET
INTELLIGENCE
RULES AND ADMINISTRATION

April 26, 2023

Mr. Satya Nadella
Chief Executive Officer
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052

Dear Mr. Nadella,

I write today regarding the need to prioritize security in the design and development of artificial intelligence (AI) systems. As companies like yours make rapid advancements in AI, we must acknowledge the security risks inherent in this technology and ensure AI development and adoption proceeds in a responsible and secure way. While public concern about the safety and security of AI has been on the rise, I know that work on AI security is not new. However, with the increasing use of AI across large swaths of our economy, and the possibility for large language models to be steadily integrated into a range of existing systems, from healthcare to finance sectors, I see an urgent need to underscore the importance of putting security at the forefront of your work. Beyond industry commitments, however, it is also clear that some level of regulation is necessary in this field.

I recognize the important work you and your colleagues are doing to advance AI. As a leading company in this emerging technology, I believe you have a responsibility to ensure that your technology products and systems are secure. I have long advocated for incorporating security-by-design, as we have found time and again that failing to consider security early in the product development lifecycle leads to more costly and less effective security. Instead, incorporating security upfront can reduce costs[1] and risks[2]. Moreover, the last five years have demonstrated that the ways in which the speed, scale, and excitement associated with new technologies have frequently obscured the shortcomings of their creators in anticipating the harmful effects of their use. AI capabilities hold enormous potential; however, we must ensure that they do not advance without appropriate safeguards and regulation.

While it is important to apply many of the same security principles we associate with traditional computing services and devices, AI presents a new set of security concerns that are distinct from

[1] Maurice Dawson et al., "Integrating Software Assurance into the Software Development Life Cycle (SDLC)," Journal of Information Systems Technology and Planning (2010), Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC
[2] "U.S. and International Partners Publish Secure-by-Design and -Default Principles and Approaches," Cybersecurity and Infrastructure Security Agency (April 13, 2023), https://www.cisa.gov/news-events/news/us-and-international-partners-publish-secure-design-and-default-principles-and-approaches

traditional software vulnerabilities. Some of the AI-specific security risks that I am concerned about include the origin, quality, and accuracy of input data (data supply chain)[3], tampering with training data (data poisoning attacks)[4], and inputs to models that intentionally cause them to make mistakes (adversarial examples)[5]. Each of these risks further highlighting the need for secure, quality data inputs. Broadly speaking, these techniques can effectively defeat or degrade the integrity, security, or performance of an AI system (including the potential confidentiality of its training data). As leading models are increasingly integrated into larger systems, often without fully mapping dependencies and downstream implications, the effects of adversarial attacks on AI systems are only magnified.

In addition to those risks, I also have concerns regarding bias, trustworthiness, and potential misuse or malicious use of AI systems. In the last six months, we have seen open source researchers repeatedly exploit a number of prominent, publicly-accessible generative models – crafting a range of clever (and often foreseeable) prompts to easily circumvent a system's rules. Examples include using widely-adopted models to generate malware[6], craft increasingly sophisticated phishing techniques[7], contribute to disinformation[8], and provide harmful information[9]. It is imperative that we address threats to not only digital security, but also threats to physical security and political security.[10]

In light of this, I am interested in learning about the measures that your company is taking to ensure the security of its AI systems. I request that you provide answers to the following questions no later than May 26, 2023.

Questions:

1. Can you provide an overview of your company's security approach or strategy?
2. What limits do you enforce on third-party access to your model and how do you actively monitor for non-compliant uses?

---

[3] "OWASP AI Security and Privacy Guide," OWASP Foundation, https://owasp.org/www-project-ai-security-and-privacy-guide/

[4] Fahri Anıl Yerlikaya, Şerif Bahtiyar, "Data poisoning attacks against machine learning algorithms", Expert Systems with Applications, Volume 208, (July 18, 2022). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417422012933

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial Examples in the Physical World," Google, Inc. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf

[6] Dan Goodin, "Hackers Are Selling A Service that Bypasses ChatGPT Restrictions on Malware," Ars Technica (February 8, 2023) https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/

[7] Lily Hay Newman, "AI Wrote Better Phishing Emails Than Humans in a Recent Test," Wired (August 7, 2021), https://www.wired.com/story/ai-phishing-emails/

[8] Tiffany Hsu, Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times (February 13, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[9] "GPT-4 Jailbreak and Hacking via RabbitHole attack, Prompt injection, Content moderation bypass and Weaponizing AI," Adversa AI (March 15, 2023), https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/

[10] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (February, 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217

3. Are you participating in third party (internal or external) test & evaluation, verification & validation of your systems?
4. What steps have you taken to ensure that you have secure and accurate data inputs and outputs? Have you provided comprehensive and accurate documentation of your training data to downstream users to allow them to evaluate whether your model is appropriate for their use?
5. Do you provide complete and accurate documentation of your model to commercial users? Which documentation standards or procedures do you rely on?
6. What kind of input sanitization techniques do you implement to ensure that your systems are not susceptible to prompt injection techniques that pose underlying system risks?
7. How are you monitoring and auditing your systems to detect and mitigate security breaches?
8. Can you explain the security measures that you take to prevent unauthorized access to your systems and models?
9. How do you protect your systems against potential breaches or cyberattacks? Do you have a plan in place to respond to a potential security incident? What is your process for alerting users that have integrated your model into downstream systems?
10. What is your process for ensuring the privacy of sensitive or personal information you that your system uses?
11. Can you describe how your company has handled past security incidents?
12. What security standards, if any, are you adhering to? Are you using NIST's AI Risk Management Framework?[11]
13. Is your company participating in the development of technical standards related to AI and AI security?
14. How are you ensuring that your company continues to be knowledgeable about evolving security best practices and risks?
15. How is your company addressing concerns about AI trustworthiness, including potential algorithmic bias and misuse or malicious use of AI?
16. Have you identified any security challenges unique to AI that you believe policymakers should address?

Thank you for your attention to these important matters and I look forward to your response.

Sincerely,

Mark R. Warner
United States Senator

[11] "AI Risk Management Framework," NIST (July 12, 2021), https://www.nist.gov/itl/ai-risk-management-framework

MARK R. WARNER
VIRGINIA

United States Senate
WASHINGTON, DC 20510–4606

COMMITTEES:
FINANCE
BANKING, HOUSING, AND URBAN AFFAIRS
BUDGET
INTELLIGENCE
RULES AND ADMINISTRATION

April 26, 2023

Mr. David Holz
Chief Executive Officer
Midjourney
611 Gateway Blvd Suite 120
South San Francisco, CA 94080

Dear Mr. Holz,

I write today regarding the need to prioritize security in the design and development of artificial intelligence (AI) systems. As companies like yours make rapid advancements in AI, we must acknowledge the security risks inherent in this technology and ensure AI development and adoption proceeds in a responsible and secure way. While public concern about the safety and security of AI has been on the rise, I know that work on AI security is not new. However, with the increasing use of AI across large swaths of our economy, and the possibility for large language models to be steadily integrated into a range of existing systems, from healthcare to finance sectors, I see an urgent need to underscore the importance of putting security at the forefront of your work. Beyond industry commitments, however, it is also clear that some level of regulation is necessary in this field.

I recognize the important work you and your colleagues are doing to advance AI. As a leading company in this emerging technology, I believe you have a responsibility to ensure that your technology products and systems are secure. I have long advocated for incorporating security-by-design, as we have found time and again that failing to consider security early in the product development lifecycle leads to more costly and less effective security. Instead, incorporating security upfront can reduce costs[1] and risks[2]. Moreover, the last five years have demonstrated that the ways in which the speed, scale, and excitement associated with new technologies have frequently obscured the shortcomings of their creators in anticipating the harmful effects of their use. AI capabilities hold enormous potential; however, we must ensure that they do not advance without appropriate safeguards and regulation.

While it is important to apply many of the same security principles we associate with traditional computing services and devices, AI presents a new set of security concerns that are distinct from

[1] Maurice Dawson et al., "Integrating Software Assurance into the Software Development Life Cycle (SDLC)," Journal of Information Systems Technology and Planning (2010), Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC
[2] "U.S. and International Partners Publish Secure-by-Design and -Default Principles and Approaches," Cybersecurity and Infrastructure Security Agency (April 13, 2023), https://www.cisa.gov/news-events/news/us-and-international-partners-publish-secure-design-and-default-principles-and-approaches

traditional software vulnerabilities. Some of the AI-specific security risks that I am concerned about include the origin, quality, and accuracy of input data (data supply chain)[3], tampering with training data (data poisoning attacks)[4], and inputs to models that intentionally cause them to make mistakes (adversarial examples)[5]. Each of these risks further highlighting the need for secure, quality data inputs. Broadly speaking, these techniques can effectively defeat or degrade the integrity, security, or performance of an AI system (including the potential confidentiality of its training data). As leading models are increasingly integrated into larger systems, often without fully mapping dependencies and downstream implications, the effects of adversarial attacks on AI systems are only magnified.

In addition to those risks, I also have concerns regarding bias, trustworthiness, and potential misuse or malicious use of AI systems. In the last six months, we have seen open source researchers repeatedly exploit a number of prominent, publicly-accessible generative models – crafting a range of clever (and often foreseeable) prompts to easily circumvent a system's rules. Examples include using widely-adopted models to generate malware[6], craft increasingly sophisticated phishing techniques[7], contribute to disinformation[8], and provide harmful information[9]. It is imperative that we address threats to not only digital security, but also threats to physical security and political security.[10]

In light of this, I am interested in learning about the measures that your company is taking to ensure the security of its AI systems. I request that you provide answers to the following questions no later than May 26, 2023.

Questions:

1. Can you provide an overview of your company's security approach or strategy?
2. What limits do you enforce on third-party access to your model and how do you actively monitor for non-compliant uses?

---

[3] "OWASP AI Security and Privacy Guide," OWASP Foundation, https://owasp.org/www-project-ai-security-and-privacy-guide/

[4] Fahri Anıl Yerlikaya, Şerif Bahtiyar, "Data poisoning attacks against machine learning algorithms", Expert Systems with Applications, Volume 208, (July 18, 2022). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417422012933

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial Examples in the Physical World," Google, Inc. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf

[6] Dan Goodin, "Hackers Are Selling A Service that Bypasses ChatGPT Restrictions on Malware," Ars Technica (February 8, 2023) https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/

[7] Lily Hay Newman, "AI Wrote Better Phishing Emails Than Humans in a Recent Test," Wired (August 7, 2021), https://www.wired.com/story/ai-phishing-emails/

[8] Tiffany Hsu, Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times (February 13, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[9] "GPT-4 Jailbreak and Hacking via RabbitHole attack, Prompt injection, Content moderation bypass and Weaponizing AI," Adversa AI (March 15, 2023), https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/

[10] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (February, 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217

3. Are you participating in third party (internal or external) test & evaluation, verification & validation of your systems?
4. What steps have you taken to ensure that you have secure and accurate data inputs and outputs? Have you provided comprehensive and accurate documentation of your training data to downstream users to allow them to evaluate whether your model is appropriate for their use?
5. Do you provide complete and accurate documentation of your model to commercial users? Which documentation standards or procedures do you rely on?
6. What kind of input sanitization techniques do you implement to ensure that your systems are not susceptible to prompt injection techniques that pose underlying system risks?
7. How are you monitoring and auditing your systems to detect and mitigate security breaches?
8. Can you explain the security measures that you take to prevent unauthorized access to your systems and models?
9. How do you protect your systems against potential breaches or cyberattacks? Do you have a plan in place to respond to a potential security incident? What is your process for alerting users that have integrated your model into downstream systems?
10. What is your process for ensuring the privacy of sensitive or personal information you that your system uses?
11. Can you describe how your company has handled past security incidents?
12. What security standards, if any, are you adhering to? Are you using NIST's AI Risk Management Framework?[11]
13. Is your company participating in the development of technical standards related to AI and AI security?
14. How are you ensuring that your company continues to be knowledgeable about evolving security best practices and risks?
15. How is your company addressing concerns about AI trustworthiness, including potential algorithmic bias and misuse or malicious use of AI?
16. Have you identified any security challenges unique to AI that you believe policymakers should address?

Thank you for your attention to these important matters and I look forward to your response.

Sincerely,

Mark R. Warner
United States Senator

[11] "AI Risk Management Framework," NIST (July 12, 2021), https://www.nist.gov/itl/ai-risk-management-framework

MARK R. WARNER
VIRGINIA

United States Senate
WASHINGTON, DC 20510–4606

COMMITTEES:
FINANCE
BANKING, HOUSING, AND URBAN AFFAIRS
BUDGET
INTELLIGENCE
RULES AND ADMINISTRATION

April 26, 2023

Mr. Sam Altman
Chief Executive Officer
OpenAI
3180 18th St
San Francisco, CA

Dear Mr. Altman,

I write today regarding the need to prioritize security in the design and development of artificial intelligence (AI) systems. As companies like yours make rapid advancements in AI, we must acknowledge the security risks inherent in this technology and ensure AI development and adoption proceeds in a responsible and secure way. While public concern about the safety and security of AI has been on the rise, I know that work on AI security is not new. However, with the increasing use of AI across large swaths of our economy, and the possibility for large language models to be steadily integrated into a range of existing systems, from healthcare to finance sectors, I see an urgent need to underscore the importance of putting security at the forefront of your work. Beyond industry commitments, however, it is also clear that some level of regulation is necessary in this field.

I recognize the important work you and your colleagues are doing to advance AI. As a leading company in this emerging technology, I believe you have a responsibility to ensure that your technology products and systems are secure. I have long advocated for incorporating security-by-design, as we have found time and again that failing to consider security early in the product development lifecycle leads to more costly and less effective security. Instead, incorporating security upfront can reduce costs[1] and risks[2]. Moreover, the last five years have demonstrated that the ways in which the speed, scale, and excitement associated with new technologies have frequently obscured the shortcomings of their creators in anticipating the harmful effects of their use. AI capabilities hold enormous potential; however, we must ensure that they do not advance without appropriate safeguards and regulation.

While it is important to apply many of the same security principles we associate with traditional computing services and devices, AI presents a new set of security concerns that are distinct from

[1] Maurice Dawson et al., "Integrating Software Assurance into the Software Development Life Cycle (SDLC)," Journal of Information Systems Technology and Planning (2010), Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC
[2] "U.S. and International Partners Publish Secure-by-Design and -Default Principles and Approaches," Cybersecurity and Infrastructure Security Agency (April 13, 2023), https://www.cisa.gov/news-events/news/us-and-international-partners-publish-secure-design-and-default-principles-and-approaches

traditional software vulnerabilities. Some of the AI-specific security risks that I am concerned about include the origin, quality, and accuracy of input data (data supply chain)[3], tampering with training data (data poisoning attacks)[4], and inputs to models that intentionally cause them to make mistakes (adversarial examples)[5]. Each of these risks further highlighting the need for secure, quality data inputs. Broadly speaking, these techniques can effectively defeat or degrade the integrity, security, or performance of an AI system (including the potential confidentiality of its training data). As leading models are increasingly integrated into larger systems, often without fully mapping dependencies and downstream implications, the effects of adversarial attacks on AI systems are only magnified.

In addition to those risks, I also have concerns regarding bias, trustworthiness, and potential misuse or malicious use of AI systems. In the last six months, we have seen open source researchers repeatedly exploit a number of prominent, publicly-accessible generative models – crafting a range of clever (and often foreseeable) prompts to easily circumvent a system's rules. Examples include using widely-adopted models to generate malware[6], craft increasingly sophisticated phishing techniques[7], contribute to disinformation[8], and provide harmful information[9]. It is imperative that we address threats to not only digital security, but also threats to physical security and political security.[10]

In light of this, I am interested in learning about the measures that your company is taking to ensure the security of its AI systems. I request that you provide answers to the following questions no later than May 26, 2023.

Questions:

1. Can you provide an overview of your company's security approach or strategy?
2. What limits do you enforce on third-party access to your model and how do you actively monitor for non-compliant uses?

---

[3] "OWASP AI Security and Privacy Guide," OWASP Foundation, https://owasp.org/www-project-ai-security-and-privacy-guide/

[4] Fahri Anıl Yerlikaya, Şerif Bahtiyar, "Data poisoning attacks against machine learning algorithms", Expert Systems with Applications, Volume 208, (July 18, 2022). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417422012933

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial Examples in the Physical World," Google, Inc. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf

[6] Dan Goodin, "Hackers Are Selling A Service that Bypasses ChatGPT Restrictions on Malware," Ars Technica (February 8, 2023) https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/

[7] Lily Hay Newman, "AI Wrote Better Phishing Emails Than Humans in a Recent Test," Wired (August 7, 2021), https://www.wired.com/story/ai-phishing-emails/

[8] Tiffany Hsu, Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times (February 13, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[9] "GPT-4 Jailbreak and Hacking via RabbitHole attack, Prompt injection, Content moderation bypass and Weaponizing AI," Adversa AI (March 15, 2023), https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/

[10] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (February, 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217

3. Are you participating in third party (internal or external) test & evaluation, verification & validation of your systems?
4. What steps have you taken to ensure that you have secure and accurate data inputs and outputs? Have you provided comprehensive and accurate documentation of your training data to downstream users to allow them to evaluate whether your model is appropriate for their use?
5. Do you provide complete and accurate documentation of your model to commercial users? Which documentation standards or procedures do you rely on?
6. What kind of input sanitization techniques do you implement to ensure that your systems are not susceptible to prompt injection techniques that pose underlying system risks?
7. How are you monitoring and auditing your systems to detect and mitigate security breaches?
8. Can you explain the security measures that you take to prevent unauthorized access to your systems and models?
9. How do you protect your systems against potential breaches or cyberattacks? Do you have a plan in place to respond to a potential security incident? What is your process for alerting users that have integrated your model into downstream systems?
10. What is your process for ensuring the privacy of sensitive or personal information you that your system uses?
11. Can you describe how your company has handled past security incidents?
12. What security standards, if any, are you adhering to? Are you using NIST's AI Risk Management Framework?[11]
13. Is your company participating in the development of technical standards related to AI and AI security?
14. How are you ensuring that your company continues to be knowledgeable about evolving security best practices and risks?
15. How is your company addressing concerns about AI trustworthiness, including potential algorithmic bias and misuse or malicious use of AI?
16. Have you identified any security challenges unique to AI that you believe policymakers should address?

Thank you for your attention to these important matters and I look forward to your response.

Sincerely,

Mark R. Warner
United States Senator

[11] "AI Risk Management Framework," NIST (July 12, 2021), https://www.nist.gov/itl/ai-risk-management-framework

MARK R. WARNER
VIRGINIA

United States Senate
WASHINGTON, DC 20510–4606

COMMITTEES:
FINANCE
BANKING, HOUSING, AND URBAN AFFAIRS
BUDGET
INTELLIGENCE
RULES AND ADMINISTRATION

April 26, 2023

Brigadier General Balan Ayyar, USAF, Retired
Chief Executive Officer
Percipient.ai
3975 Freedom Cir. Suite 850
Santa Clara, CA 95054

Dear Brigadier General Ayyar,

I write today regarding the need to prioritize security in the design and development of artificial intelligence (AI) systems. As companies like yours make rapid advancements in AI, we must acknowledge the security risks inherent in this technology and ensure AI development and adoption proceeds in a responsible and secure way. While public concern about the safety and security of AI has been on the rise, I know that work on AI security is not new. However, with the increasing use of AI across large swaths of our economy, and the possibility for large language models to be steadily integrated into a range of existing systems, from healthcare to finance sectors, I see an urgent need to underscore the importance of putting security at the forefront of your work. Beyond industry commitments, however, it is also clear that some level of regulation is necessary in this field.

I recognize the important work you and your colleagues are doing to advance AI. As a leading company in this emerging technology, I believe you have a responsibility to ensure that your technology products and systems are secure. I have long advocated for incorporating security-by-design, as we have found time and again that failing to consider security early in the product development lifecycle leads to more costly and less effective security. Instead, incorporating security upfront can reduce costs[1] and risks[2]. Moreover, the last five years have demonstrated that the ways in which the speed, scale, and excitement associated with new technologies have frequently obscured the shortcomings of their creators in anticipating the harmful effects of their use. AI capabilities hold enormous potential; however, we must ensure that they do not advance without appropriate safeguards and regulation.

While it is important to apply many of the same security principles we associate with traditional computing services and devices, AI presents a new set of security concerns that are distinct from

[1] Maurice Dawson et al., "Integrating Software Assurance into the Software Development Life Cycle (SDLC)," Journal of Information Systems Technology and Planning (2010), Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC
[2] "U.S. and International Partners Publish Secure-by-Design and -Default Principles and Approaches," Cybersecurity and Infrastructure Security Agency (April 13, 2023), https://www.cisa.gov/news-events/news/us-and-international-partners-publish-secure-design-and-default-principles-and-approaches

traditional software vulnerabilities. Some of the AI-specific security risks that I am concerned about include the origin, quality, and accuracy of input data (data supply chain)[3], tampering with training data (data poisoning attacks)[4], and inputs to models that intentionally cause them to make mistakes (adversarial examples)[5]. Each of these risks further highlighting the need for secure, quality data inputs. Broadly speaking, these techniques can effectively defeat or degrade the integrity, security, or performance of an AI system (including the potential confidentiality of its training data). As leading models are increasingly integrated into larger systems, often without fully mapping dependencies and downstream implications, the effects of adversarial attacks on AI systems are only magnified.

In addition to those risks, I also have concerns regarding bias, trustworthiness, and potential misuse or malicious use of AI systems. In the last six months, we have seen open source researchers repeatedly exploit a number of prominent, publicly-accessible generative models – crafting a range of clever (and often foreseeable) prompts to easily circumvent a system's rules. Examples include using widely-adopted models to generate malware[6], craft increasingly sophisticated phishing techniques[7], contribute to disinformation[8], and provide harmful information[9]. It is imperative that we address threats to not only digital security, but also threats to physical security and political security.[10]

In light of this, I am interested in learning about the measures that your company is taking to ensure the security of its AI systems. I request that you provide answers to the following questions no later than May 26, 2023.

Questions:

1. Can you provide an overview of your company's security approach or strategy?
2. What limits do you enforce on third-party access to your model and how do you actively monitor for non-compliant uses?

---

[3] "OWASP AI Security and Privacy Guide," OWASP Foundation, https://owasp.org/www-project-ai-security-and-privacy-guide/

[4] Fahri Anıl Yerlikaya, Şerif Bahtiyar, "Data poisoning attacks against machine learning algorithms", Expert Systems with Applications, Volume 208, (July 18, 2022). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417422012933

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial Examples in the Physical World," Google, Inc. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf

[6] Dan Goodin, "Hackers Are Selling A Service that Bypasses ChatGPT Restrictions on Malware," Ars Technica (February 8, 2023) https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/

[7] Lily Hay Newman, "AI Wrote Better Phishing Emails Than Humans in a Recent Test," Wired (August 7, 2021), https://www.wired.com/story/ai-phishing-emails/

[8] Tiffany Hsu, Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times (February 13, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[9] "GPT-4 Jailbreak and Hacking via RabbitHole attack, Prompt injection, Content moderation bypass and Weaponizing AI," Adversa AI (March 15, 2023), https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/

[10] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (February, 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217

3. Are you participating in third party (internal or external) test & evaluation, verification & validation of your systems?
4. What steps have you taken to ensure that you have secure and accurate data inputs and outputs? Have you provided comprehensive and accurate documentation of your training data to downstream users to allow them to evaluate whether your model is appropriate for their use?
5. Do you provide complete and accurate documentation of your model to commercial users? Which documentation standards or procedures do you rely on?
6. What kind of input sanitization techniques do you implement to ensure that your systems are not susceptible to prompt injection techniques that pose underlying system risks?
7. How are you monitoring and auditing your systems to detect and mitigate security breaches?
8. Can you explain the security measures that you take to prevent unauthorized access to your systems and models?
9. How do you protect your systems against potential breaches or cyberattacks? Do you have a plan in place to respond to a potential security incident? What is your process for alerting users that have integrated your model into downstream systems?
10. What is your process for ensuring the privacy of sensitive or personal information you that your system uses?
11. Can you describe how your company has handled past security incidents?
12. What security standards, if any, are you adhering to? Are you using NIST's AI Risk Management Framework?[11]
13. Is your company participating in the development of technical standards related to AI and AI security?
14. How are you ensuring that your company continues to be knowledgeable about evolving security best practices and risks?
15. How is your company addressing concerns about AI trustworthiness, including potential algorithmic bias and misuse or malicious use of AI?
16. Have you identified any security challenges unique to AI that you believe policymakers should address?

Thank you for your attention to these important matters and I look forward to your response.

Sincerely,

Mark R. Warner
United States Senator

[11] "AI Risk Management Framework," NIST (July 12, 2021), https://www.nist.gov/itl/ai-risk-management-framework

MARK R. WARNER
VIRGINIA

United States Senate
WASHINGTON, DC 20510–4606

COMMITTEES:
FINANCE
BANKING, HOUSING, AND URBAN AFFAIRS
BUDGET
INTELLIGENCE
RULES AND ADMINISTRATION

April 26, 2023

Mr. Alexandr Wang
Chief Executive Officer
Scale AI
155 5th St
San Francisco, CA 94103

Dear Mr. Wang,

I write today regarding the need to prioritize security in the design and development of artificial intelligence (AI) systems. As companies like yours make rapid advancements in AI, we must acknowledge the security risks inherent in this technology and ensure AI development and adoption proceeds in a responsible and secure way. While public concern about the safety and security of AI has been on the rise, I know that work on AI security is not new. However, with the increasing use of AI across large swaths of our economy, and the possibility for large language models to be steadily integrated into a range of existing systems, from healthcare to finance sectors, I see an urgent need to underscore the importance of putting security at the forefront of your work. Beyond industry commitments, however, it is also clear that some level of regulation is necessary in this field.

I recognize the important work you and your colleagues are doing to advance AI. As a leading company in this emerging technology, I believe you have a responsibility to ensure that your technology products and systems are secure. I have long advocated for incorporating security-by-design, as we have found time and again that failing to consider security early in the product development lifecycle leads to more costly and less effective security. Instead, incorporating security upfront can reduce costs[1] and risks[2]. Moreover, the last five years have demonstrated that the ways in which the speed, scale, and excitement associated with new technologies have frequently obscured the shortcomings of their creators in anticipating the harmful effects of their use. AI capabilities hold enormous potential; however, we must ensure that they do not advance without appropriate safeguards and regulation.

While it is important to apply many of the same security principles we associate with traditional computing services and devices, AI presents a new set of security concerns that are distinct from

[1] Maurice Dawson et al., "Integrating Software Assurance into the Software Development Life Cycle (SDLC)," Journal of Information Systems Technology and Planning (2010), Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC
[2] "U.S. and International Partners Publish Secure-by-Design and -Default Principles and Approaches," Cybersecurity and Infrastructure Security Agency (April 13, 2023), https://www.cisa.gov/news-events/news/us-and-international-partners-publish-secure-design-and-default-principles-and-approaches

traditional software vulnerabilities. Some of the AI-specific security risks that I am concerned about include the origin, quality, and accuracy of input data (data supply chain)[3], tampering with training data (data poisoning attacks)[4], and inputs to models that intentionally cause them to make mistakes (adversarial examples)[5]. Each of these risks further highlighting the need for secure, quality data inputs. Broadly speaking, these techniques can effectively defeat or degrade the integrity, security, or performance of an AI system (including the potential confidentiality of its training data). As leading models are increasingly integrated into larger systems, often without fully mapping dependencies and downstream implications, the effects of adversarial attacks on AI systems are only magnified.

In addition to those risks, I also have concerns regarding bias, trustworthiness, and potential misuse or malicious use of AI systems. In the last six months, we have seen open source researchers repeatedly exploit a number of prominent, publicly-accessible generative models – crafting a range of clever (and often foreseeable) prompts to easily circumvent a system's rules. Examples include using widely-adopted models to generate malware[6], craft increasingly sophisticated phishing techniques[7], contribute to disinformation[8], and provide harmful information[9]. It is imperative that we address threats to not only digital security, but also threats to physical security and political security.[10]

In light of this, I am interested in learning about the measures that your company is taking to ensure the security of its AI systems. I request that you provide answers to the following questions no later than May 26, 2023.

Questions:

1. Can you provide an overview of your company's security approach or strategy?
2. What limits do you enforce on third-party access to your model and how do you actively monitor for non-compliant uses?

---

[3] "OWASP AI Security and Privacy Guide," OWASP Foundation, https://owasp.org/www-project-ai-security-and-privacy-guide/

[4] Fahri Anıl Yerlikaya, Şerif Bahtiyar, "Data poisoning attacks against machine learning algorithms", Expert Systems with Applications, Volume 208, (July 18, 2022). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417422012933

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial Examples in the Physical World," Google, Inc. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf

[6] Dan Goodin, "Hackers Are Selling A Service that Bypasses ChatGPT Restrictions on Malware," Ars Technica (February 8, 2023) https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/

[7] Lily Hay Newman, "AI Wrote Better Phishing Emails Than Humans in a Recent Test," Wired (August 7, 2021), https://www.wired.com/story/ai-phishing-emails/

[8] Tiffany Hsu, Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times (February 13, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[9] "GPT-4 Jailbreak and Hacking via RabbitHole attack, Prompt injection, Content moderation bypass and Weaponizing AI," Adversa AI (March 15, 2023), https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/

[10] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (February, 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217

3. Are you participating in third party (internal or external) test & evaluation, verification & validation of your systems?
4. What steps have you taken to ensure that you have secure and accurate data inputs and outputs? Have you provided comprehensive and accurate documentation of your training data to downstream users to allow them to evaluate whether your model is appropriate for their use?
5. Do you provide complete and accurate documentation of your model to commercial users? Which documentation standards or procedures do you rely on?
6. What kind of input sanitization techniques do you implement to ensure that your systems are not susceptible to prompt injection techniques that pose underlying system risks?
7. How are you monitoring and auditing your systems to detect and mitigate security breaches?
8. Can you explain the security measures that you take to prevent unauthorized access to your systems and models?
9. How do you protect your systems against potential breaches or cyberattacks? Do you have a plan in place to respond to a potential security incident? What is your process for alerting users that have integrated your model into downstream systems?
10. What is your process for ensuring the privacy of sensitive or personal information you that your system uses?
11. Can you describe how your company has handled past security incidents?
12. What security standards, if any, are you adhering to? Are you using NIST's AI Risk Management Framework?[11]
13. Is your company participating in the development of technical standards related to AI and AI security?
14. How are you ensuring that your company continues to be knowledgeable about evolving security best practices and risks?
15. How is your company addressing concerns about AI trustworthiness, including potential algorithmic bias and misuse or malicious use of AI?
16. Have you identified any security challenges unique to AI that you believe policymakers should address?

Thank you for your attention to these important matters and I look forward to your response.

Sincerely,

Mark R. Warner
United States Senator

[11] "AI Risk Management Framework," NIST (July 12, 2021), https://www.nist.gov/itl/ai-risk-management-framework

MARK R. WARNER
VIRGINIA

United States Senate
WASHINGTON, DC 20510–4606

COMMITTEES:
FINANCE
BANKING, HOUSING, AND URBAN AFFAIRS
BUDGET
INTELLIGENCE
RULES AND ADMINISTRATION

April 26, 2023

Mr. Emad Mostaque
Chief Executive Officer
Stability AI
88 Notting Hill Gate
London, England, W11 3HP

Dear Mr. Mostaque,

I write today regarding the need to prioritize security in the design and development of artificial intelligence (AI) systems. As companies like yours make rapid advancements in AI, we must acknowledge the security risks inherent in this technology and ensure AI development and adoption proceeds in a responsible and secure way. While public concern about the safety and security of AI has been on the rise, I know that work on AI security is not new. However, with the increasing use of AI across large swaths of our economy, and the possibility for large language models to be steadily integrated into a range of existing systems, from healthcare to finance sectors, I see an urgent need to underscore the importance of putting security at the forefront of your work. Beyond industry commitments, however, it is also clear that some level of regulation is necessary in this field.

I recognize the important work you and your colleagues are doing to advance AI. As a leading company in this emerging technology, I believe you have a responsibility to ensure that your technology products and systems are secure. I have long advocated for incorporating security-by-design, as we have found time and again that failing to consider security early in the product development lifecycle leads to more costly and less effective security. Instead, incorporating security upfront can reduce costs[1] and risks[2]. Moreover, the last five years have demonstrated that the ways in which the speed, scale, and excitement associated with new technologies have frequently obscured the shortcomings of their creators in anticipating the harmful effects of their use. AI capabilities hold enormous potential; however, we must ensure that they do not advance without appropriate safeguards and regulation.

While it is important to apply many of the same security principles we associate with traditional computing services and devices, AI presents a new set of security concerns that are distinct from

[1] Maurice Dawson et al., "Integrating Software Assurance into the Software Development Life Cycle (SDLC)," Journal of Information Systems Technology and Planning (2010), Available at: https://www.researchgate.net/publication/255965523_Integrating_Software_Assurance_into_the_Software_Development_Life_Cycle_SDLC
[2] "U.S. and International Partners Publish Secure-by-Design and -Default Principles and Approaches," Cybersecurity and Infrastructure Security Agency (April 13, 2023), https://www.cisa.gov/news-events/news/us-and-international-partners-publish-secure-design-and-default-principles-and-approaches

traditional software vulnerabilities. Some of the AI-specific security risks that I am concerned about include the origin, quality, and accuracy of input data (data supply chain)[3], tampering with training data (data poisoning attacks)[4], and inputs to models that intentionally cause them to make mistakes (adversarial examples)[5]. Each of these risks further highlighting the need for secure, quality data inputs. Broadly speaking, these techniques can effectively defeat or degrade the integrity, security, or performance of an AI system (including the potential confidentiality of its training data). As leading models are increasingly integrated into larger systems, often without fully mapping dependencies and downstream implications, the effects of adversarial attacks on AI systems are only magnified.

In addition to those risks, I also have concerns regarding bias, trustworthiness, and potential misuse or malicious use of AI systems. In the last six months, we have seen open source researchers repeatedly exploit a number of prominent, publicly-accessible generative models – crafting a range of clever (and often foreseeable) prompts to easily circumvent a system's rules. Examples include using widely-adopted models to generate malware[6], craft increasingly sophisticated phishing techniques[7], contribute to disinformation[8], and provide harmful information[9]. It is imperative that we address threats to not only digital security, but also threats to physical security and political security.[10]

In light of this, I am interested in learning about the measures that your company is taking to ensure the security of its AI systems. I request that you provide answers to the following questions no later than May 26, 2023.

Questions:

1. Can you provide an overview of your company's security approach or strategy?
2. What limits do you enforce on third-party access to your model and how do you actively monitor for non-compliant uses?

---

[3] "OWASP AI Security and Privacy Guide," OWASP Foundation, https://owasp.org/www-project-ai-security-and-privacy-guide/

[4] Fahri Anıl Yerlikaya, Şerif Bahtiyar, "Data poisoning attacks against machine learning algorithms", Expert Systems with Applications, Volume 208, (July 18, 2022). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417422012933

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio, "Adversarial Examples in the Physical World," Google, Inc. Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf

[6] Dan Goodin, "Hackers Are Selling A Service that Bypasses ChatGPT Restrictions on Malware," Ars Technica (February 8, 2023) https://arstechnica.com/information-technology/2023/02/now-open-fee-based-telegram-service-that-uses-chatgpt-to-generate-malware/

[7] Lily Hay Newman, "AI Wrote Better Phishing Emails Than Humans in a Recent Test," Wired (August 7, 2021), https://www.wired.com/story/ai-phishing-emails/

[8] Tiffany Hsu, Stuart A. Thompson, "Disinformation Researchers Raise Alarms About A.I. Chatbots," New York Times (February 13, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[9] "GPT-4 Jailbreak and Hacking via RabbitHole attack, Prompt injection, Content moderation bypass and Weaponizing AI," Adversa AI (March 15, 2023), https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/

[10] Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," (February, 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217

3. Are you participating in third party (internal or external) test & evaluation, verification & validation of your systems?
4. What steps have you taken to ensure that you have secure and accurate data inputs and outputs? Have you provided comprehensive and accurate documentation of your training data to downstream users to allow them to evaluate whether your model is appropriate for their use?
5. Do you provide complete and accurate documentation of your model to commercial users? Which documentation standards or procedures do you rely on?
6. What kind of input sanitization techniques do you implement to ensure that your systems are not susceptible to prompt injection techniques that pose underlying system risks?
7. How are you monitoring and auditing your systems to detect and mitigate security breaches?
8. Can you explain the security measures that you take to prevent unauthorized access to your systems and models?
9. How do you protect your systems against potential breaches or cyberattacks? Do you have a plan in place to respond to a potential security incident? What is your process for alerting users that have integrated your model into downstream systems?
10. What is your process for ensuring the privacy of sensitive or personal information you that your system uses?
11. Can you describe how your company has handled past security incidents?
12. What security standards, if any, are you adhering to? Are you using NIST's AI Risk Management Framework?[11]
13. Is your company participating in the development of technical standards related to AI and AI security?
14. How are you ensuring that your company continues to be knowledgeable about evolving security best practices and risks?
15. How is your company addressing concerns about AI trustworthiness, including potential algorithmic bias and misuse or malicious use of AI?
16. Have you identified any security challenges unique to AI that you believe policymakers should address?

Thank you for your attention to these important matters and I look forward to your response.

Sincerely,

Mark R. Warner
United States Senator

[11] "AI Risk Management Framework," NIST (July 12, 2021), https://www.nist.gov/itl/ai-risk-management-framework