UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC. COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: Case No. 1:23-cv-11195 Case No. 1:24-cv-03285 Case No. 1:24-cv-04872	Case No. 1:25-md-3143-SHS-OTW

DECLARATION OF SARAH BIRD IN SUPPORT OF MICROSOFT CORPORATION'S MOTION FOR SUMMARY JUDGMENT

I, Sarah Bird, hereby declare as follows:

A. Background

1. I am the Chief Product Officer of Responsible AI at Microsoft and submit this declaration in support of Microsoft's Motion for Summary Judgment. Unless stated otherwise, all facts stated herein are within my personal knowledge. If called upon, I would be willing to testify as set forth in this declaration.

2. In February 2026, my deposition was taken in this case as Microsoft's designated witness under Rule 30(b)(6) concerning numerous topics related to Microsoft's Responsible AI efforts. My deposition was taken again in August 2026 as a non-retained expert witness concerning substantially related topics.

3. Before joining Microsoft, I earned a Ph.D. in Computer Science and a Master of Science in Computer Science from the University of California, Berkeley.

4. I initially joined Microsoft in 2013 as a post-doctoral researcher and subsequently held a position as a Technical Advisor focused on research related to the

integration of machine learning into emerging technology systems. From 2017 to 2019, I was a Technical Program Manager at Meta (formerly Facebook) where I was a founding member of the company's AI ethics group.

5. In my current position as Microsoft's Chief Product Officer of Responsible AI since February 2019, I lead Microsoft's cross-company efforts to integrate responsible AI principles and practices in the development and deployment of AI products. During my career at Microsoft, I have been centrally involved with efforts to develop Microsoft's organizational Responsible AI program, practices, and initiatives.

a. For instance, I have been an active member of Microsoft's AI Ethics and Effects in Engineering and Research ("AETHER") Committee over many years, a research initiative dedicated to responsible AI efforts focused on fairness and transparency.

b. I was also a founding researcher of Microsoft's Fairness, Accountability, Transparency, and Ethics in AI ("FATE") group, which studies the societal implications posed by AI and aims to facilitate the innovation of computational techniques that can be integrated into responsible AI systems.

c. I later co-founded Microsoft's Responsible AI Group, Microsoft's governance body that sets company-wide Responsible AI principles and technical standards to operationalize their implementation.

6. In my capacity as an engineer, I was also centrally involved with the development and deployment of major Microsoft AI products, including GitHub Copilot and Microsoft Copilot (formerly known as Bing Chat), and specifically worked on building responsible AI technologies into the foundational system designs. Relatedly, I have led

the development of responsible AI tools, including Fairlearn, SmartNoise, InterpretML, ASSERT, and Agent Control Specification.

7. Additionally, I have had extensive engagement with external stakeholders to advance both Microsoft's and the broader AI industry's adoption of responsible AI practices that align with the evolving needs of customers and society at large.

a. For instance, I, on behalf of Microsoft, collaborate with external experts, researchers, organizations, and partners (*e.g.*, OpenAI and Anthropic) to develop state-of-the-art responsible AI practices that are consistently implemented across the industry.

b. Additionally, I frequently participate in public speaking engagements, including conferences and interviews, on such topics.

c. I have also facilitated such opportunities for cross-industry engagement, including co-founding the MLSys Research Conference (tailored toward research regarding best practices, design principles, and development of learning methods for machine learning systems) and Learning Systems Workshops (a working group for machine learning experts to explore challenges when building machine-learning systems and develop implementable solutions).

8. I have also authored numerous scholarly publications spanning several areas of research, such as hardware architecture, operating systems, and large-scale computing systems, applied machine learning systems, and operationalization at scale. My more recent publications are focused on responsible AI, including measurable risks, mitigation algorithms, and engineering solutions for the deployment of compliant AI

systems, which collectively address the multi-faceted array of the responsible AI framework.

B. Development of Microsoft's Responsible AI Practices and Efforts

9. Fundamentally, Responsible AI ("RAI") is the discipline of designing, building, and deploying AI systems so that they are safe and appropriate for their intended uses, while minimizing foreseeable harms.

10. At Microsoft, RAI is both a practice and an organizational culture. It has developed into a comprehensive framework of principles, corporate governance, engineering practices, public transparency, policy advocacy, and product-specific safeguards. Microsoft strives to support the RAI ecosystem through large-scale internal and external investments to drive broad RAI adoption and innovation.

a. Microsoft's Internal Responsible AI Governance

11. The foundation of Microsoft's internal RAI efforts relies on strong governance and coordination at scale to ensure the end-to-end development and deployment of its AI product portfolio in a manner that is safe, fair, transparent, privacy-protective, accountable, and aligned with human values.

12. From my experience engaging with the broader AI industry, Microsoft is an early RAI adopter and industry leader. From my industry observations, Microsoft is one of the earliest large technology companies to formalize RAI governance and operationalize RAI practices, which it began well before the generative AI boom. Beginning in 2017, Microsoft established the AETHER Committee to research the ethical implications of AI, which informed its first set of Responsible AI Principles adopted in 2018. Also in 2018, Microsoft established the first dedicated AI red team in the industry, a group of interdisciplinary experts dedicated to probing the responsible AI risks

associated with AI systems. Thereafter, Microsoft's RAI program expanded through the establishment of its Office of Responsible AI ("ORA") in 2019, which was formed to develop a comprehensive approach to RAI governance and oversight by setting formal company-wide policies and putting those policies into practice. Microsoft's adoption of its second version of Responsible AI Standards followed in 2022.

13. Additionally, Microsoft and OpenAI engaged in early collaborations to establish formal requirements to ensure the safe deployment of AI systems according to the companies' shared longstanding commitment to RAI practices. In anticipation of the broader adoption and deployment of OpenAI's advanced LLM technology, the companies established the Joint Deployment Safety Board (the "Joint DSB") in 2021, which defined a comprehensive pre-release safety review process informed by the second version of Microsoft's Responsible AI Standards. The Joint DSB conducted safety reviews ahead of the first deployment of an OpenAI large language model with highly advanced capabilities, such as GPT-4.

14. Microsoft has continued to cement RAI into its core business practices. Today, Microsoft's RAI program relies on a cross-organizational governance system to ensure the continued oversight, guidance, and progress of responsible AI policies and practices. This institutional framework relies on a multi-tiered structure involving executive-level leadership from Microsoft's Board of Directors; senior-level leadership from the Responsible AI Council; establishment of policies and practices by the Office of Responsible AI, including product deployment oversight from its Deployment Safety Board (the "DSB" or "Microsoft's DSB"); and contributions from Microsoft Research, Policy, and Engineering teams to drive innovation and scale responsible AI practices throughout the company.

15. From my observations of the larger industry, Microsoft takes a leading approach with respect to its cross-company approach to operationalizing RAI principles and practices. This is in part driven by Microsoft's unique position in the industry—it releases thousands of AI-powered products and features annually, including general purpose consumer AI products, general and industry-specific workplace AI products, highly-customized AI applications for enterprise customers, developer AI tools, and AI builder platforms. Due to the volume and diversity of Microsoft's AI product suite, coordination at scale is necessary to ensure that the development and deployment of all such systems is in alignment with Microsoft's RAI Standards.

b. Industry and Customer Engagement

16. Beyond its internal operations, Microsoft also invests substantial efforts and resources in facilitating coordination at scale between stakeholders to advance the state-of-the-art of RAI principles, standards, and practices.

17. Microsoft's core RAI practices include regular external engagement to understand the various needs of different stakeholders, promote transparency, and reach consensus-based RAI solutions. This dialogue informs Microsoft's independent standards and commitments, including its Responsible AI Standards, AI Customer Commitments, Customer Copyright Commitment, AI Access Principles, and Transparency Reports. Additionally, Microsoft has joined in commitments to governance and regulatory frameworks. Some of the more prominent examples include the White House Voluntary AI Commitment, NIST AI Risk Management Framework, EU AI Act Compliance, G7 AI Governance Framework (Hiroshima AI Process), and Seoul Summit Frontier AI Safety Commitments. Microsoft has not only joined in such commitments, but it has also

published its Blueprint for Governing AI, a formal policy proposal directed at regulators that was designed to capture stakeholder consensus on RAI best practices.

18. Further, Microsoft has led efforts to establish industry bodies and standards dedicated to responsible AI development and deployment. For instance, Microsoft is a founding member of the Frontier Model Forum, C2PA (Coalition for Content Provenance and Authenticity), Partnership on AI, Hiroshima AI Process Reporting Framework, MLCommons, UNESCO AI Business Council, Global Perspectives Responsible AI Fellowship, amongst others.

19. Microsoft views exchanging knowledge and facilitating access to technical resources as critically important to the widespread adoption and advancement of RAI standards and practices.

a. One of Microsoft's early key contributions included making its initially internal Responsible AI Standard publicly available. Microsoft recognized the importance of sharing its foundational knowledge from which the industry could continue to build upon, inviting feedback and contributing to the discussion about building better norms and practices around AI.

b. Microsoft subsequently committed to regularly publishing Transparency Reports, which serve the dual purposes of promoting accountability as well as informing the AI community and customers of Microsoft's progress, learnings, and ongoing efforts in its Responsible AI Program.[1]

[1] Responsible-AI-Transparency-Report-2024.pdf, *Microsoft*, 2024; Responsible-AI-Transparency-Report-2025-vertical.pdf, *Microsoft*, 2025; 2026 Responsible AI Transparency Report, *Microsoft*, 2026.

c. Microsoft's continuous efforts also include collaborative efforts between its own research teams and external industry researchers, practitioners, and experts to understand the risks and implications posed by AI and work to develop practical and implementable RAI solutions.

20. Finally, Microsoft makes responsible AI technology available to empower its customers to develop AI systems and applications responsibly. These span governance tools, model assessment and measurement tools, generative AI safety tools, security tools, and open-source RAI tool libraries.

C. Microsoft's Responsible AI Standards, Principles, and Practices

21. Microsoft's RAI framework is guided by a principled and actionable approach to ensure that its novel AI systems are developed and deployed responsibly. In other words, Microsoft's AI systems are designed to maximize their benefits to society while minimizing potential risks.

22. This framework is grounded in six core principles that govern and guide Microsoft's AI development and deployment: fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. Microsoft's Responsible AI Standard codifies these principles into concrete and actionable engineering requirements under a "defense-in-depth" approach.

23. Microsoft uses an expert-driven approach to implement responsible practices end-to-end in the AI development lifecycle using a three-stage process: (1) mapping potential risks; (2) measuring risks using evaluations and testing; and (3) managing risks through mitigations and controls. Importantly, Microsoft's Responsible AI Standard is designed as a living document. Each standard is written broadly without

limitation to a defined set of risks such that it remains durable as this fast-moving technology continues to evolve and new risks emerge.

a. "Mapping" refers to identifying the risks associated with the AI system. This begins with a holistic assessment of the AI system's intended use cases and undesirable outcomes. Microsoft, in collaboration with a network of experts, conducts extensive testing to understand how risks might manifest in practice and inform early mitigation strategies. Beyond testing the system when used as ordinarily expected, red teaming or stress testing is a key mechanism used to intentionally elicit undesirable model behaviors, which facilitates a deeper understanding of the limitations and vulnerabilities of the system.

b. Once risks are identified, the "measuring" phase focuses on systematically evaluating identified risks to inform how to manage such risks and refine the system's behavior. Microsoft uses defined metrics to measure any identified risks that also serve as a baseline for mitigations performance testing. These metrics are neither static nor uniformly applied; rather, metrics may be updated to account for new risks as they are identified, and the acceptable thresholds for measurable risk may be adjusted depending on the system's intended use case.

c. "Risk management" is an ongoing, multifaceted approach. During the development phase, risk management involves integration of techniques and tools into the AI system itself at either (or both) the model and application levels to address known or potential risks at run time. After system deployment, Microsoft uses ongoing monitoring and oversight

channels to ensure that the system is operating as intended. When previously unknown risks or unexpected user behaviors are identified, pre-established processes for incident response are employed. Additionally, Microsoft has learned that a phased release strategy, beginning with a controlled release to a limited number of trusted users followed by additional incremental releases, helps to map, measure, and manage risks that emerge in practice ahead of broader release.

24. Importantly, responsible AI is not an effort with a finish line, but a continuing responsibility that persists throughout the entire lifecycle of an AI system as technology and user behavior evolve. Accordingly, Microsoft's RAI practices are cyclical and sustained efforts to ensure that deployed AI systems continue to adhere to Microsoft's foundational RAI principles and safeguards adapt to emerging societal implications.

D. Responsible Development and Deployment of Copilot

25. Microsoft designed, developed, and deployed Copilot (formerly known as Bing Chat) in accordance with the above-described defense-in-depth protocol. Microsoft employed the full suite of RAI tools and techniques to mitigate potential risks while also securing the beneficial and transformative uses of the technology.

a. Map, Measure, and Mitigate Framework Applied to Copilot

26. Copilot's intended uses informed Microsoft's approach to responsibly developing and deploying the system. Copilot was designed to help people learn and discover new information and be more productive and creative while safeguarding both customers and society against potential risks that may surface with the release of this novel generative AI technology.

27. The "mapping" stage began at the model level. OpenAI's GPT-4 large language model powered Copilot in its first iteration. As this was a highly advanced model with unprecedented capabilities, GPT-4 underwent Microsoft's and OpenAI's rigorous Joint DSB Safety Review process. The companies collaborated to evaluate the model's standalone core capabilities and limitations, including extensive red teaming designed to produce harmful responses and surface potential avenues for misuse. This process informed the companies' understanding of the model's potential vulnerabilities and risks at the broadest level and likewise advised early mitigation strategies for subsequent integration into the Copilot system.

28. In addition to the model-level assessments, Microsoft conducted extensive rounds of testing at the application level, *i.e.*, the model as integrated into Microsoft's Copilot generative AI system, focused specifically on its web search and conversational capabilities. A multidisciplinary team of experts conducted additional red team testing as well as probed potential risks that could arise in benign scenarios. After launch, Microsoft continued its testing and evaluations in collaboration with expert teams, using new learnings to improve and expand its approach on an ongoing basis.

29. While isolated red-team examples can surface instances of specific harms, real users engage in millions of unique, multi-turn conversations. To address this challenge, Microsoft proactively developed new RAI metrics tailored to Copilot's conversational chat experience, *e.g.*, jailbreaks, harmful content generation, bias, ungrounded content, etc. Recognizing that this advanced AI system required advanced techniques to ensure its responsible deployment, Microsoft developed a partially automated measurement pipeline to evaluate these risks at scale.

30. After identifying and assessing the scope of potential risks, a range of RAI safeguards were implemented:

- **Post-Training:** To manage GPT-4's potential RAI risks, OpenAI conducted post-training to guide intended model behavior, such as improving instruction following and grounding capabilities, as well as steering the model away from unintended behaviors, *e.g.*, hallucinations or harmful content generation.
- **Classifiers:** An ensemble of classifiers are used to help detect user prompts or model responses that contain certain categories of content associated with RAI risks (*e.g.*, violence, hate, sexual, offensive, or self-harm) and detect user abuses or misuses of the system (*e.g.*, jailbreak attempts). If detected, this classifier-based safety system may filter content, change the topic, refuse to generate content, or redirect the user to traditional search results.
- **Metaprompting (or "Prompt Instructions"):** Microsoft designed "metaprompt" instructions provided to the model and every time it generates a user prompt. Metaprompts are used to guide intended model behavior in alignment with Microsoft's RAI Principles and user expectations, as well as to steer the model away from unintended behaviors across the full spectrum of known RAI risks. Because the LLMs powering are post-trained to optimize their instruction-following capabilities, metaprompting is an especially effective tool that can be used to address the full range of known RAI risks.

- **Grounding:** For certain conversations where users are seeking information, Copilot's retrieval-augmented generation (or "grounding") capabilities are used to provide the model with information sourced from Bing web search results. Because Bing search is designed to retrieve relevant, highly ranked, and credible webpages, this has the downstream effect of ensuring that Copilot responses are based on accurate and reliable sources of information. This is a primary defense mechanism for preventing hallucinations or disseminating misinformation.
- **User-Interface Features:** If a Copilot answer is grounded in web search results, links to webpages are included along with visible disclosures that users are interacting with AI. These features are intended to help users use the system appropriately and avoid overreliance on Copilot-generated outputs. The interface also includes mechanisms for user feedback, which helps Microsoft monitor product performance and address issues as they may arise.

31. Additionally, a phased release strategy is a core approach for responsible AI system deployment. As such, Microsoft initially launched Copilot in limited preview, followed by additional incremental releases. This allowed Microsoft to observe real-world use, identify emerging issues, and improve safeguards to ensure that the system operated as intended before it was made widely accessible.

32. Importantly, LLM behavior can vary dramatically on a model-by-model basis. Models likewise may exhibit varying levels of risk across the RAI taxonomy that require different types and degrees of intervention, and they also may respond differently to RAI mitigations and interventions. Each new model deployment (including both

frontier models and post-trained versions of previously released models) thus requires repeating the development cycle, notwithstanding that the model is integrated into the same fundamental AI system.

33. Microsoft's commitment to responsible deployment did not end with Copilot's release; responsible deployment is an iterative, continuous cycle employed with the same careful attention throughout the product's life. Post-deployment monitoring and oversight are critical mechanisms for ensuring that Copilot continues to operate in adherence with Microsoft's Responsible AI Principles and that its Responsible AI Standards evolve alongside the technology.

a. Responsible AI Safeguards for Copyrighted Content Generation

34. Microsoft designed Copilot to harness its language generation capabilities to support users in engaging with conversational experiences to explore information across a wide range of topics, promote free and open access to information, spark creativity by helping users create content (*e.g.*, poems, jokes, stories, letters), enhance productivity, amongst many other applications having transformative potential.

35. However, these capabilities are accompanied by an inherent risk that the LLM that powers Copilot—if not subject to mitigation—may generate outputs that include some portion of copyrighted material. Consistent with its Responsible AI Principles of Reliability, Accountability and Fairness, Microsoft has built safeguards into Copilot to address the risk of this unintended outcome, whether user-elicited or resulting from a system defect.

36. There are three ways in which the chatbot may generate outputs containing verbatim excerpts of copyrighted content: regurgitation, grounding content reproduction,

and hallucination. While all three are distinct system behaviors, as explained below, they each result in the same unintended, undesirable outcome that Microsoft addresses collectively.

a. "Regurgitation" occurs when the LLM generates outputs that reproduce verbatim or near-verbatim portions of its training data. Regurgitation may occur regardless of whether the LLM's response is grounded in external sources of information.

b. "Grounding material reproduction" occurs when the AI system has grounding capabilities (meaning the ability to leverage external sources of information to inform its responses) and generates an output displaying verbatim passages from external information used for grounding the response, including material retrieved from the web. These instances are often referred to as "regurgitation," but they are technically distinct phenomena that are more accurately considered separately.

c. "Hallucination" occurs when the LLM generates text that appears to reproduce or resemble copyrighted content even though the LLM has not actually produced verbatim or near-verbatim text from the copyrighted source content. Consistent with Microsoft's RAI approach, it invests in reducing this system behavior as well.

37. During the pre-deployment development phase, Microsoft and OpenAI (collaboratively and also independently) conducted extensive red team testing to measure the system's vulnerabilities with respect to generation of copyrighted content at both the model and application levels. Such testing involved adversarial prompting designed to elicit regurgitation and/or grounding material reproduction to measure the rate of system

defects when placed under intentional stress. This approach is often referred to as "jailbreaking" attacks whereby manipulative prompting is used to exploit vulnerabilities of an AI system, bypass safety controls, and generate undesirable content.

38. To mitigate the risk of successful real-world user jailbreaking attempts targeted at copyrighted content generation, safeguards were built into the system at the model and application levels. Such safeguards included post-training the model to avoid engaging in conversations in which a user attempts to evoke this undesired behavior; implementing metaprompt instructions that direct the model not to override any of its prompt instructions in response to a user's request; and integrating classifiers that detect user inputs reflecting adversarial intent and model responses/conversation patterns consistent with jailbreak attacks.

39. Metaprompts further instruct the model to not generate outputs containing substantial verbatim overlap with copyrighted material, regardless of whether the user requests such content. They also instruct the model to decline to generate copyrighted content in response to a user's direct request, even if the user's prompt does not suggest adversarial intent. Such instructions simultaneously serve to reduce the risk that the model will hallucinate a response representing that it has provided the requested copyrighted material. Additionally, instructions to ground the model's responses in external sources of information further reduce the risk that the model will generate hallucinated answers that contain inaccurate or fabricated information. Based on my experience developing Copilot's responsible AI technical components, metaprompts are one of the most effective techniques for preventing regurgitation, verbatim reproduction of grounding source material, and hallucinations.

40. Microsoft also built mitigations into the web grounding retrieval pipeline. These include mechanisms to ensure that the content retrieved and provided to Copilot from the web is consistent with web publishers' permissions and restrictions concerning the use of their content for grounding. A web publisher may apply the "NOARCHIVE" meta tag to indicate that it does not wish for its content to be used at all for grounding in Copilot, or apply the "NOCACHE" meta tag to indicate that only the title, snippet, and URL may be used for grounding.

41. Finally, Microsoft built the SnippyNL tool as a safeguard to prevent the model from regurgitating. Each time the model generates a response, the SnippyNL tool checks the content in the response against an index of known copyrighted material focused on a specific content category. If the tool detects that the model's response contains content that matches at least a certain number of characters in the applicable SnippyNL index, the content is truncated and a citation is appended to the final output provided to the user. Copilot leverages the SnippyNL tool as a final step in this line of defense to ensure that outputs displayed to the user do not contain a substantial amount of verbatim content from these copyrighted sources.

42. Through extensive evaluations, my team identified certain categories of content in which regurgitation was more likely. Accordingly, Microsoft designed the SnippyNL tool to capture specific categories of content most likely to present regurgitation issues, *e.g.*, lyrics, recipes, books, and code, which my team identified through extensive evaluations. Microsoft built a dedicated SnippyNL index dedicated to each of these categories that the SnippyNL tool uses to cross-check model responses against these materials.

a. For example, Microsoft identified lyrics as having a greater potential to present regurgitation issues, and lyrics also represent content that users are reasonably expected to request Copilot to generate.

b. By contrast, the data did not reveal that news content presented the same concern. Users requesting news-related content from a model generally seek current information that post-dates model training, so the likelihood of regurgitation is extremely slim.

c. Instead, model responses containing verbatim news content were more likely due to defects in the web grounding pipeline. Accordingly, Microsoft addresses verbatim reproduction of news content through robust protections built into the retrieval system itself, in addition to the safeguards described above, which aim to eliminate any substantial verbatim reproduction of grounding data.

43. As with all of Microsoft's RAI tools, SnippyNL's design was (and is) continuously improved to minimize regurgitation risks while maximizing Copilot's societal benefits.

a. Specifically, Microsoft has engaged in an iterative experimental process to determine a maximum character threshold for verbatim content matches that most accurately captures this undesired system behavior, using a data-driven approach to maximize precision and recall.

b. "Precision" measures the proportion of content identified as presenting a particular RAI risk that actually presents that risk. "Recall" measures the proportion of all instances of that risk that the detection system successfully identifies. High precision reduces false positives (*i.e.*, avoids blocking

benign content), whereas high recall reduces false negatives (*i.e.*, avoids missing behaviors presenting the risk).

c. The system presents an inherent tradeoff between precision and recall, as improving one metric often reduces the other. My experience developing the SnippyNL tool revealed that high recall can result in over-detection (which Microsoft developers have also referred to as "over-blocking"), meaning that the tool would detect and block responses that used generic statements rather than capture meaningful copyrighted content. While verbatim reproduction of copyrighted material is an undesired system behavior, over-detection presents a harm to the competing societal interests of fostering free and open access to information, avoiding censorship, and addressing fairness concerns with erasure.

d. While there is no perfect numeric threshold for preventing every instance in which the system output includes a meaningful amount of verbatim copyrighted content, Microsoft's approach reflects a solution designed to balance the rights of content creators with both the rights and benefits to users, meanwhile remaining flexible to adapt to evolving needs and new learnings.

I declare under penalty of perjury that the foregoing is true and correct.

Executed on September 02, 2026.

Signed by:
Sarah Bird
E9AE40507D5142D...

Sarah Bird