Date: October 29, 2023

BRIA Artificial Intelligence Ltd.
Vered Horesh
Chief Strategic AI Partnerships
vered@bria.ai
126 Yigal Alon St.
Tel Aviv 6744332
Israel

To: The United States Copyright Office

Re: Response to Notice of Inquiry - Artificial Intelligence and Copyright

Submission from BRIA Artificial Intelligence Ltd., leading developer of responsible AI enterprise solutions trained exclusively on licensed data from esteemed partners like Getty Images, Alamy of the PA Media Group, and Envato[1].

Introduction

BRIA AI is an Israeli artificial intelligence company founded in 2020 with the mission to develop an ethical visual generative AI creative platform suitable for commercial use. The company is backed by prominent venture capital firms from the USA, Israel, Korea and Japan. Led by experts in computer vision, machine learning and computer science from top universities, our team has pioneered generative models and products, following stringent ethical standards. We are proud to have earned the trust of enterprises and partners worldwide by ensuring our AI systems meet the highest standards of transparency, accountability, robustness, and safety.

At BRIA AI, we knew solving for ethical AI requires rigorous data acquisition practices. To that end, we created the world's largest multi-source visual training dataset under contract. Among our esteemed data partners are Getty Images, Alamy, Envato and many more agencies, photographers, artists and illustrators.

BRIA AI stands apart as one of very few technology companies worldwide that develops full-stack generative AI systems by training diffusion models entirely from scratch. Through this ambitious undertaking, BRIA AI has acquired extensive unique expertise across the entire AI development pipeline - from curating training datasets to architecting models to managing responsible deployments. This ground-up

[1] The perspectives expressed in this submission represent the independent views of BRIA AI alone, and do not necessarily reflect or represent the opinions or endorsements of our data partners

approach provides BRIA AI unparalleled technical insight into the inner workings, capabilities, and risks of generative models based on hands-on experience.

As both an innovator developing pioneering AI technologies, and a responsible industry steward committed to demystifying these complex systems, BRIA AI is uniquely positioned to provide authoritative perspectives to inform the Copyright Office's inquiry. We believe our on-the-ground expertise training generative models from the ground up offers critical clarity on the nuanced copyright implications of emerging AI - both the risks requiring diligent oversight, and the opportunities for creativity and access that thoughtful policy can unlock. Our goal is to demonstrate an ethical approach to AI that respects content creators' copyright is not only viable for a for-profit company but also ensures a sustainable ecosystem.

BRIA AI's selected data partners

Summary of Position

Artificial intelligence promises to transform entire industries and enhance human creativity in unprecedented ways. But as this rapidly evolving technology continues to advance, it poses complex questions for intellectual property systems worldwide. As an AI company committed to responsible innovation that benefits society, BRIA AI welcomes the Copyright Office's timely inquiry on Artificial Intelligence and Copyright. We aim to contribute constructive perspectives to this crucial debate over balancing creativity, commerce, and the public interest. In the following paragraphs, we share our vision for how AI copyright policy can incentivize equitable growth across the AI generation ecosystem, from data sources to end users. With thoughtful guidelines, we can unleash AI's creative potential while upholding the Constitutional aims of copyright – "To promote the Progress of Science and useful Arts."

Our philosophy at BRIA AI is that the development of meaningful generative AI technologies requires fair compensation for all efforts involved. While it is industry

standard to compensate for AI expertise and computing power, data acquisition practices for training generative AI models commonly involve indiscriminately harvesting vast amounts of content from the internet and other repositories, ignoring creators' rights.

BRIA AI developed a proprietary data attribution technology (patent pending) as described in further detail below. This unique technology connects supply and demand, maintaining incentives to create quality works that are in demand, rather than compensation programs that provide equal incentives regardless of market needs. It aligns the interests of BRIA AI and our data partners, transforming original works into a recurring revenue stream that is fair and transparent. This technology and our unique business model earned us the trust and collaboration of our data partners.

We also created the Fair Diffusion program where top illustrators can set their own pricing for AI art generation inspired by their style and works. These artists joined because of our unparalleled ethical approach to generative AI.

Artists and content creators do not fear progression, only unfair compensation. Like compute vendors and AI experts receive, creators deserve fair payment for their creative contributions to innovative technologies.

But ethical AI means more things to us at BRIA AI: opt-in only, opt-out anytime, and restricting synthetic outputs that violate societal norms. We adhere to these principles.

How to create a sustainable ecosystem from creators to end users? In addition to compensating data sources, certain AI outputs must warrant copyright protection. Copyright protection should incentivize participation across the ecosystem. Copyright for AI users will drive adoption, knowing they can monetize valuable outputs. Wider legal use will enable AI companies to profit from greater demand for their solutions, justifying both R&D investments and creators' fair compensation. Standards around transparency and attribution allow all parties - creators, technology companies, and AI tools users - to properly claim economic benefits.

This balanced approach aligns incentives for all stakeholders. Attribution of outputs copyright to users, with acknowledgment of data sources, encourages participation. Customers are incentivized to legally leverage AI. AI companies profit from transparent practices. And original creators receive due credit and compensation. This will maximize innovation and economic benefits for all.

AI has already disrupted markets for human creation, and this impact will grow. Sustaining incentives for human creativity in the age of AI carries cultural and societal importance. Our commitment to a sustainable ecosystem will allow authentic human expression to continue evolving in tandem with AI, preventing outdated output, homogenization of creative output and diminished output. Recent studies support these concerns around AI trained on synthetic data[2]. We believe our approach charting a sustainable path forward for human and artificial creativity together.

Another important aspect of knowing every item of data used to train our models is unparalleled safety and transparency. Our dataset is commercial grade, so our models have not been exposed to unsafe, harmful, unauthorized public figure content or privacy infringing content. This makes it extremely difficult to generate unsafe, unethical, or misinformative synthetic output. We closely monitor all our data for fairness and non-discrimination. As a result, the generated output tends to be less biased. The provenance of our training data enables safety and fairness by design in our AI models, rather than trying to filter unacceptable outputs after the fact.

BRIA AI's output safety

[2] ***The Curse of Recursion: Training on Generated Data Makes Models Forget***, Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, Ross Anderson

Response to Clause 7 of the Copyright Office's Inquiry: The Process of Training AI Models

Training Diffusion Models

Text-to-image models are trained to synthesize high-quality images according to a textual description, often called a prompt. Training of text-to-image AI models begins with curating a large dataset of image-text pairs. The training data is divided into batches or subsets that are sequentially fed into the model architecture. This involves permanent copies of images along their captions in a static storage as well as making exact digital copies of the data temporarily stored in memory and memorized in the models parameters.

The model is usually tasked with repairing a noisy/damaged image with the guidance of the textual prompt. For example, diffusion models are trained to remove noise from a visually noisy image conditioned on the prompt. This way the model learns the relationship between textual descriptions and visual images. The training data batches propagate through the neural network, updating the model parameters (by memorizing the original input image-text pairs) to improve its ability to repair the images based on the text. This propagation-update cycle repeats for many epochs until the model achieves the target performance.

Overall, model training involves extensive digital reproduction of the training dataset to update the model's internal parameters. Although theoretically training data could be discarded from the storage by the end of the training, in practice most companies maintain archives of the specific training datasets. There are several reasons for retaining training data:

1. To retrain or fine-tune models on updated datasets for accuracy improvements;
2. To enable traceability and monitoring of training sources for transparency; and/or
3. To support internal research into model behaviors and errors to guide further development.

In addition to that, by the end of the training, the model itself holds in its parameters a memorization of some of the input image-caption relations pairs (although those are not trivial to extract).

In practice, full datasets are commonly retained post-training by companies for a variety of purposes. This persistent archiving means the reproduction of works

during training can convert to much longer-term copies within AI development ecosystems. The retention duration varies based on use cases but can span multiple years to support retraining, research, and other audit purposes.

Inference of Diffusion Models

At inference time, the trained model can then generate new images from pure noise based on text prompts only, having established the cross-modal connections between language and image features.

The model parameters retain the generalized relationships learned during training but are naturally capable of memorizing images from the training set to optimize their training objective. In some cases, artifacts of specific training data may get embedded in the parameters. This could result in memorization and reconstruction of unique inputs. Indeed, recent research works have shown that some training images are memorized[3]. In other cases, the images are not memorized, but the model learns a general concept. A classic example of memorization is asking the model to produce the "Mona Lisa by Leonardo da Vinci"; of course, potent models would accurately reproduce the original piece.

Predicting which images will be memorized and which will not is an unsolved problem currently, and the same holds for controlling which images will be memorized. Therefore, we conclude that generating memorized images is inevitable.

When the model encounters a novel prompt during the inference phase, it generates a response based on the patterns and relationships it has internalized during training. Current research demonstrates that such models encapsulate hierarchical structures of semantic concepts, where some concepts are built on top of other more specific concepts[4]. For example, the "US President" encapsulates the representation of different presidents like Donald Trump or Barack Obama. Usually, this does not involve reproducing specific training examples, but rather applying the

[3] ***Extracting Training Data from Diffusion Models***, Carlini, Nicolas and Hayes, Jamie and Nasr, Milad and Jagielski, Matthew and Sehwag, Vikash and Tramer, Florian and Balle, Borja and Ippolito, Daphne and Wallace, Eric

Understanding and Mitigating Copying in Diffusion Models, Somepalli, Gowthami and Singla, Vasu and Goldblum, Micah and Geiping, Jonas and Goldstein, Tom

[4] **The Hidden Language of Diffusion Models**, Chefer, Hila and Lang, Oran and Geva, Mor and Polosukhin, Volodymyr and Shocher, Assaf and Irani, Michal and Mosseri, Inbar and Wolf, Lior

generalized concepts the model has learned. However, as mentioned above, there are instances where the model may reproduce memorized images from the training set.

Can Diffusion Models Forget?

The concept of "unlearning" or "forgetting" specific inferences in an AI model has been a subject of ongoing research. While some studies have demonstrated the feasibility of this process, it is typically limited to erasing a single or a small number of concepts from the model's knowledge base.

One of the inherent challenges in unlearning is that concepts within AI models are intricately intertwined. This means that attempting to remove or alter one concept can inadvertently affect other related or even seemingly unrelated concepts. For instance, if the model were to unlearn information about specific individuals like "Barack Obama" or "Donald Trump," this process could potentially distort the broader concept of "US president," as these individual instances are deeply interconnected within the model's parameter space.

Furthermore, considering the vast number of concepts an AI model learns during training, particularly those which might violate copyright laws, unlearning at a large scale becomes impractical. Attempting to do so would likely result in significant degradation of the model's performance and utility, as countless interrelated concepts would be affected.

From an economic standpoint, the process of unlearning is not feasible when considering the scale of retraining or modification required to achieve it without harming the model's integrity. The computational resources, time, and associated costs make it an impractical solution, especially given the current state of technology.

While techniques for directly "unlearning" problematic inferences in AI models remain limited, other complementary mitigation approaches are gaining traction for managing risks post-development. Watermarking training data, controlling model deployment, and monitoring outputs take a different tack - not attempting to surgically alter models but instead tracking, containing, and addressing issues downstream through enhanced transparency, attribution, and oversight. These methods do not solve the root cause, given unlearning's feasibility barriers. Rather, they provide policy guardrails to deter and respond to harms once models are already trained and deployed. As the field grapples with the challenges of selective forgetting, bolstering these downstream mitigations that enhance accountability offers a pragmatic path forward.

Watermarking involves subtly altering select training examples with identifiable signals to discern if those samples get embedded in the model's parameters or outputs. Watermarking aids auditing and attribution but does not directly enable unlearning. It provides signal tracking for monitoring, rather than removing, problematic data traces.

Carefully controlling how trained models are deployed, queried, and shared publicly can mitigate risks from unwanted inferences or memorization, without resorting to unlearning. Limiting access to core model APIs and parameters maintains oversight, unlike releasing open-source code and data which enables uncontrolled cloning.

Actively monitoring the model's outputs using human review, classifiers, and other testing can identify problematic generations like copyright violations or unfair biases. While not erasing the source, active monitoring combined with takedowns provides oversight and management of issues.

While these supplementary techniques help strengthen oversight and attribution, they do not resolve the core need to mitigate problematic data from being ingested in the first place. This underscores the urgency of rigorous dataset curation as an essential line of defense, given unlearning's barriers. Proactive vetting and filtering of training data provides a foundation for AI integrity from the start, rather than relying solely on after-the-fact monitoring and control of deployments. It aligns with the adage that "an ounce of prevention is worth a pound of cure." By investing in "data hygiene", the industry can nurture healthier AI from the ground up, reducing reliance on imperfect attempts to surgical harm. This ethos is reflected in the growing emphasis on and progress around responsible data practices.

Leading industry and academia AI research groups blend automation cues with nuanced human discernment, scaled through crowdsourcing, demonstrating a responsible approach to deep data curation, despite the rapidly growing volumes involved in AI development. This focus on "data hygiene" stems from recognizing that clean data upstream prevents having to rely on complex unlearning procedures downstream. Major firms are also funding academic efforts to develop scalable tools and frameworks for responsible data curation, benefiting the broader ecosystem. High profile model failures have further motivated the rapid adoption of robust training data vetting. While significant challenges remain in scalable curation, the progress highlights that rigorous data practices are increasingly seen as an essential first line of defense among industry and academia leaders seeking to mitigate AI risks. The rise of startups focused on data curation and large investments by incumbents

reflect the growing momentum and priority of this mitigation technique within the field.

In summary, while selective unlearning remains an active research area, it faces substantial barriers to feasibility as a solution, especially for large commercial state-of-the-art AI models. Given these challenges, complementing unlearning efforts with transparency, monitoring, and deployment controls provides pragmatic policy guardrails. However, rigorous training data curation stands out as a critical preventative line of defense, enabling mitigation by design rather than as an after-the-fact corrective. The momentum growing around responsible data practices reflects a recognition that nurturing integrity early in the AI lifecycle reduces reliance on imperfect attempts to surgically undo harm. While advances in selective unlearning are worthwhile to pursue, instilling "data hygiene" promises greater impact for fostering the development of ethical, accountable AI systems.

Identifying Training Data Usage in Absence of Access

In most cases, definitively determining whether specific training data was used is extremely difficult without direct access to the dataset.

Models have complex inner representations distributed across parameters, so explicit artifacts of specific training data are rarely visible. In addition, the training process of AI models is inherently stochastic. This means that the way in which a specific image or piece of data influences the model can vary widely between different training runs. In one instance, the model might end up memorizing the image, while in another, it might completely overlook it, with a spectrum of possibilities in between. At the same time, equal outputs could often emerge from various different training sources. Lack of access to the actual data severely limits auditing techniques based on model output alone.

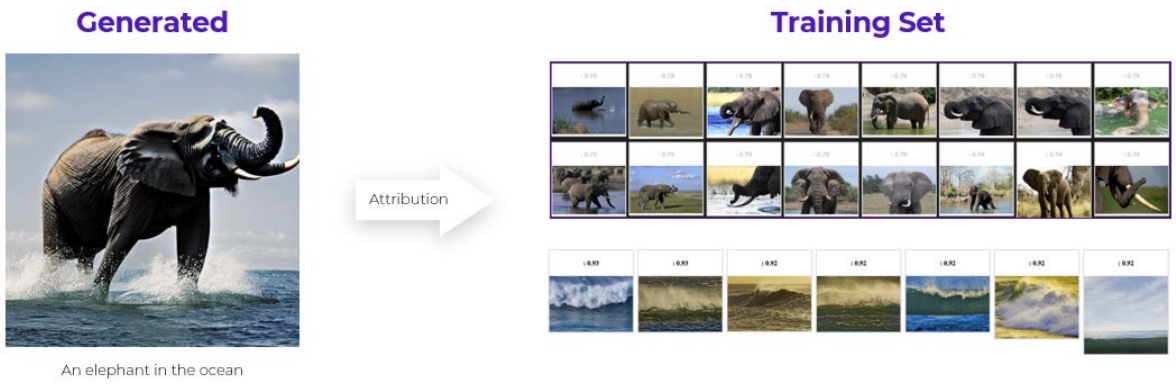
Some techniques like training data watermarking could provide indications if those embedded signals manifest in model outputs, as described above. Detailed analysis of model behavior on a wide corpus of inputs might uncover patterns indirectly suggesting imprints of specific data. However, current methods lack robustness and precision for unambiguous attribution absent access to the underlying training dataset. There are also significant scalability challenges to analyzing the massive parameter spaces of large state-of-the-art models and the exponentially vast set of potential training sources.

Access to model parameters and architecture details could assist analysis, but still faces fundamental obstacles. Reverse engineering training data from models alone, without any access to the original datasets, remains an extremely challenging if not impossible task given current techniques. In addition, having full visibility and access to the model's internal parameters and workings, goes beyond the typical API-level access provided to users. Given that AI models are considered valuable intellectual property, providing such deep access to external parties poses significant challenges, as it could jeopardize the protection and security of this intellectual property.

In summary, while an area of active research, unambiguously auditing AI models for use of specific training data without any access to the actual underlying dataset remains extremely difficult if not impossible currently.

Response to Clause 12 of the Copyright Office's Inquiry: Identifying A Particular Work's Contribution to Generative AI Output

BRIA AI developed proprietary attribution technology (patent pending), that identifies the original works within our training dataset that most impacted each synthetic output and allocates revenue from that output to the creators of the most impactful works.



BRIA AI's attribution mechanism

At BRIA AI, we have created an attribution mechanism aimed at transparent and fair distribution of value among authentic source works. While the complex nature of diffusion models means attribution may not have absolute scientific precision, our technology provides sufficient approximation to be considered equitable. By factoring the relative contributions of training sources, our methodology distributes attribution and rewards in a way conceived as fair and acceptable across creators. Though inherently approximate, our approach brings unprecedented

transparency to AI value distribution compared to treating generative works as wholly novel creations unmoored from source material context.

BRIA AI takes a major step forward from today's widespread neglect and appropriation of source training material. Progress should not be obstructed by waiting for flawless scientific attribution. Reasonably good faith efforts based on multilateral signals can distribute value fairly, even if the underlying creative relationships evade absolute codification. We believe this collaborative path forward is preferable for both AI developers and data contributors.

Closing Notes: The Path Forward - Fostering Continued Progress through Collaboration

As AI capabilities continue rapidly advancing, there are understandable concerns and uncertainties about impacts on existing creative ecosystems. However, we believe there is a constructive path forward based on transparency, fair compensation, and cross-industry collaboration.

BRIA AI aims to demonstrate this through our rigorous data licensing, attribution technologies, and partnerships bridging data providers, AI developers, and commercial users. We have only scratched the surface of AI's vast potential to enhance human creativity, productivity, and knowledge. But realizing this potential requires inclusive policy making bringing together stakeholders across the value chain.

The Copyright Office's inquiry exemplifies this collaborative spirit, and we appreciate the opportunity to contribute BRIA AI's insights stemming from real-world experience training generative models. We believe strongly in AI's role in improving lives. But human values must remain at the center - nurturing empowerment, accountability, and distributed opportunity. We commit to enabling human creativity to thrive alongside AI creativity in a sustainable, equitable future.

Sincerely yours,

Vered Horesh, Chief Strategic AI Partnerships
BRIA Artificial Intelligence Ltd.