Hannah Ismael*

Examining generative image models amidst privacy regulations

https://doi.org/10.1515/jigs-2024-0009
Published online November 27, 2024

Abstract: As diffusion models emerge as a new frontier in generative AI, requiring vast image databases as their inputs, the question arises: how should regulators approach policies concerning the collection and utilization of these images? Though generative image models currently interpret the data they scrape as public, regulatory bodies have yet to confirm this as a viable understanding. This paper explores the current public/personal distinction of data as well as the respective legal standards for both categories in both the American and European context. This paper acts as a guide for regulators seeking to understand monopolization and privacy implications of confirming the validity of using open sourced images versus imagining a reality of curated or licensed datasets amidst outrage from artists over a breach of an expectation of collection/use to their artwork. Though arguments have been made regarding using copyright to protect artists, this paper seeks to explore other pathways for regulating generative image models under our current conceptual frameworks of privacy.

Keywords: tech policy; GDPR; privacy; monopolization; antitrust

1 Background

Stability AI released Stable Diffusion, a text to image model that allows users to prompt an algorithm to create a work emulating a particular artist or including certain visual elements. Stable Diffusion, as well as other emerging generative image models such as Midjourney and Dalle2, were trained on a segment of LAION 5B, a dataset of text image pairs compiled through web crawling across the internet (What is the Super Huge Data Set "LAION-5B" That Made a Great Contribution to the Development of Image Generation AI "Stable Diffusion"? 2022). LAION has evaded privacy regulations by being not a dataset of images itself, but rather, an index of links to images that can be used for model training purposes. As a result, LAION is able to push the burden of accessing data to whichever entity is seeking to web-scrape; in this instance, these entities are image generative models. Though the recent LinkedIn v. hiQ ruling has established precedent for legalizing web-scraping publicly available data within a certain proprietary context, there has not been precedent established for web-scraping such a large public domain to be monetized later down the pipeline. This claim, that the data scraped through web crawling, is truly public information is brought to question in this paper ("HiQ Labs v. LinkedIn" 2022). The ambiguity has prompted a slew of lawsuits, with both individual artists and companies, such as Getty Images, outraged over the use of their images to train these models, and calls for reimagining regulations in light of the new technologies that have arisen.

2 Methodology

This paper is a meta analysis review of existing literature, legal sources, and empirical case studies. The primary case study this paper focuses on is the Stable Diffusion data scraping scandal, which was chosen given that its case inspired the questions at the heart of this paper. The topics explored (public/private distinction, legal bases/rights, monopolization concerns) flowed as a natural logical extension of the primary case study. Auxiliary case studies chosen had to speak to the issue of personal/public distinction, an inconsistency with the legality of its justification for use (legal bases), issues with legal rights not preserved, or monopolization concerns that could result from such models. I chose case studies the most relevant case studies that most clearly articulated each relevant section or highlighted a point of contention with another case study.

The literature referenced were selected based on their contributions towards different conceptions of data privacy and competition policy. I chose literature that succinctly exemplified the tension between the two incentives, and fit within the discussion of generative image models. Literature also included news articles of the primary case study (the LAION 5B data scraping/Stable Diffusion incident), which helped explain the fundamental perspectives of different stakeholders. Legal sources were used to explain the existing

***Corresponding author: Hannah Ismael**, UC Berkeley, 101 Sproul Hall, Berkeley, CA, USA, E-mail: hannahismael42@berkeley.edu

privacy regulations, such as the CCPA, GDPR, and a number of non-legislative informal regulations made through FTC rulings.

3 Public or personal data?

Stable Diffusion has claimed their images are open source, which implies these images are not personal data. Their claim is not yet supported or opposed by regulatory decision makers or legal precedent. Several privacy frameworks are available for analyzing the legality of gathering imaging data. These frameworks consider a range of factors, including whether this data qualifies as artistic intellectual property or falls under the category of proprietary data (such as trade secrets), which might have the potential to compromise the business interests of companies. This paper will focus on the public/personal distinction as conceptualized under the California Consumer Protection Act (CCPA) and the EU's General Data Protection Regulation (GDPR), since companies like Stability AI are conducting their business as if the data they scraped fell under public domain within these schemas. en

Under GDPR, personal data is any data that is "related to an identified or identifiable natural person"("Art. 4 GDPR – Definitions," 2016). This means that data is considered personal if it can be used to identify an individual, even if there is not a name attached to it. Different pieces of an individual's seemingly unrelated data are considered personal if, in conglomeration, they can be used to identify said individual. Similarly, under the CCPA, personal information is data that "identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household" (California Consumer Privacy Act 2020).

The input dataset for Stable Diffusion's model is the dataset from LAION 5B, which contains 5.8 billion image-text pairs and can be considered personal data under CCPA and GDPR. The LAION 5B dataset focuses on photos it ranks as "aesthetic" based on an aesthetic predictive score (What Is the Super Huge Data Set "LAION-5B" 2022). Frequent source domains of these aesthetic images include Pinterest, Shopify, Redbubble, Etsy, Wikimedia, DeviantArt, and stock image sites. The text-image pairs have enabled an indexing into an unprecedented quantity of images, and the ease with which these images can be accessed raises concerns given that there is not yet conclusive legal precedent regarding linked or indexed data. Images linked within the dataset arguably fall under GDPR and CCPA provisions as personally identifiable: Pinterest images and pictures of models from clothing websites undoubtedly contain faces, and an artist's work on their online store may have their signature or be in a style so distinct that it is identifiably them (Baio 2022).

For the text-image pairs in the LAION 5B dataset, the potential for identifiability exists within a spectrum, illustrating the limitations of binary data privacy classification. Some artists are blatantly identifiable with their names attached in the text, as has been the case for artists Karla Ortiz and Greg Rutkowski (Hill 2023), who have joined the class action lawsuit against Stability AI. For artists whose identifiability is their style, the potential for identifiability is more ambiguous.

Although there are still considerable ambiguities to be resolved in the public/private distinction of information, the usefulness of this binary distinction on the basis of identification may be limited due to technological advances in de-anonymization methods. Some privacy scholars have claimed that the framework of personal information, in being based on identification in both GDPR and CCPA statutes, appears increasingly inconsequential as computer scientists find that anonymized data is easily de-anonymized, and data in conjunction with other data can be used to easily identify an individual (Tene and Polonetsky 2013). In a study examining the identifiability of anonymous authors, Narayanan et al. (2012) were able to re-identify anonymous authors solely based on the author's style. Though anonymized data in the past has allowed businesses the ability to pursue ventures while still maintaining some semblance of respecting privacy, it's become clear as the complexity of data collection has increased, anonymization/de-identification is not a permanent condition (Purtova 2018). In order to make such a point, one privacy scholar, Purtova, conducted a case study including weather information collected from a weather station in the Dutch city of Eindhoven. Because the weather station, which alone collects rainfall, wind speed, wind direction, and temperature, can be tied to other city crime prevention data such as audio sensors and video footage, Purtova explains how individuals can be identified from such data. She argues that weather, a seemingly innocuous piece of information, has become personal. Purtova concluded "irreversible anonymity is no longer possible" (Purtova 2018).

Furthermore, the personal nature of these images may be expanded simply through the added dimension of information that these images present–their city/location, and their interests. However, it is increasingly challenging to distinguish personal data using this binary, as the rather expansive datification of life would mean almost all information could be deemed personal, thereby nullifying public/private distinction's power. A GDPR provision does claim that identifiability remains dependent on the reasonability of an individual being identified, effectively steering the

conversation towards context dependence in determining the public-personal binary. However, "as the data processing technologies advance, and the pool of data which can be combined grows … so does the reasonable likelihood of somebody being able to link any piece of information to a person" (Purtova 2018). As Tene and Polonetsky have explained, a more useful framework for explaining the harm and risk potential of identifying an individual could be in the form of a risk assessment matrix or some other continuous taxonomy (Tene and Polonetsky 2013). In the case of the images in the LAION 5B dataset, a privacy framework with different thresholds depending on the severity of harm if identified could be a useful conceptualization.

4 Legal bases and rights dependent on PII distinction

In spite of the fact that increasing ease of identifiability is blurring the distinction between public and personal data, data protection laws such as GDPR and CCPA that rely on this distinction still act as the primary legal foundation for prosecutors seeking to establish red lines for data controllers. There are different legal bases for data collection and legal rights for data owners under the CCPA and GDPR depending on whether data is classified as personal or public. In the EU, GDPR differentiates between personal and public data for legal bases of data collection, and a company must have a legal basis for processing data if it is categorized as personal. However, under US law, no such legal basis is necessary to process personal data, though there is an expectation that data collection is not "unfair or deceptive" if the data is collected from consumers. Under both the CCPA and GDPR, data deemed personal entitles data subjects to certain legal rights. Though scholars have acknowledged the public/personal distinction between data has become rather dated, it is helpful to examine the legal bases and legal rights still dependent on this classification, especially in the case that the public/personal schema remains. In addition, it's important to examine where a web scraped data source such as the LAION 5B one would fall amidst existing privacy regulation, as privacy authorities have yet to decide where the data collection falls amidst this framework.

4.1 FTC section 5 as an alternative to substantive privacy law

Under both CCPA and GDPR, a substantial amount of legislative power hinges on the personal classification of the data in question. As a result, companies working toward generative modeling solutions often strive to meet the criteria for public data in order to circumvent these legislative requirements. In these cases, it may be necessary to find new legal bases for prosecuting companies that meet the criteria for public data yet present substantial threats to the well-being of the data subjects.

If one was to assume these images are all public information, as Stable Diffusion claims, then the dataset is legal in its collection. No legal basis is required under the CCPA, regardless of personal data status or not, and the GDPR only requires a legal basis for collecting data if the data is personal. No such legal rights exist for data not deemed personal; in other words, public data does not inherit legal rights. Consequently, there is minimal legislative justification for regulating practices that may harm society, but do not fall neatly into the personal data distinction.

4.1.1 FTC regulations

The FTC's power in being able to regulate unfair, deceptive, or monopolistic practices that can hurt consumer welfare allows the FTC to act as a privacy body by examining practices through the consumer protection lens. This means that, though many of the cases they take concern personal data, data being personal is not a necessity for them to seek legal remedy. In certain cases where personal data is concerned and California courts, using the CCPA, could have litigated against, the FTC may have been used as the enforcement agency of choice due to their larger bureaucratic capacity and the lack of needing to adhere to such a distinction.

One of the U.S.'s most powerful regulations in commerce is Section 5A of the FTC Act, which "bars unfair and deceptive acts and practices in or affecting commerce." Conventionally used against false advertising and deceptive marketing, this section has increasingly been used as the FTC's legal backing to act as an informal regulatory body against data privacy abuses in the past 10 years. For the FTC to successfully levy Section 5 of the FTC Act, they must prove the company has committed unfair or deceptive business practices. The FTC must prove how the company or individual in question met the legal standards for either unfair or deceptive business practices.

A well-known case where the FTC used Section 5 was the Cambridge Analytica Scandal, where Cambridge Analytica was fined for deceptively collecting the data of millions of users without their consent for voter profiling and subsequent political ad targeting ("FTC Issues Opinion and Order" 2019). Section 5 was successfully linked to user expectations; customers did not expect to have their data from a personality quiz initiated by Cambridge Analytica used for political advertising purposes nor did they expect to have their

associated Facebook User ID used. Despite promising to not collect their name, Cambridge Analytica harvested their User ID (which effectively identifies an individual). However, it's important to note that in this case, the consumer (those that Cambridge Analytica is advertising to) and those whom Cambridge Analytica is collecting data from are the same. In this manner, Section 5, a consumer protection act, was used to protect individuals from extensive data collection practices that have caused harm through identifying these individuals as consumers.

Should the FTC decide to use Section 5 of the FTC Act against generative image models, they could claim that the data collection of image data was a deceptive practice. The agency could explain that these companies acted in a deceptive manner by portraying the datasets they collected from as truly public sources, when privacy standards for proprietary and copyright law in regards to personal data are still being developed.

Despite the FTC's rulings against unfair or deceptive collections of data, the FTC itself is not an agency devoted to protecting an individual's privacy, at least not unless said individual is a customer. The case of generative AI is a unique one: those from whom the data is scraped (artists) are not the same as the targeted customer (art buyers). Thus, for the FTC to weigh in on such a case, they would need to shift towards being a body that regulated privacy in a more general sense. This may be possible, as the FTC has taken a clear stance against data broking and behavioral targeting with data collection. However, regulating through the FTC would inherently mean that the analysis of privacy would be done through the lens of a business interaction and the welfare of the consumer. In cases where the consumer-provider relationship is not so clearly defined, such as that of generative image models, this could be a chance for the FTC to pivot, or a chance to reinforce the US interpretation of privacy being a component of competition.

4.2 Personal data, legal bases

There are no separate rules for obtaining a legal basis for personal data besides what has been outlined for public data under the CCPA. However, under the GDPR, a data "controller," or entity responsible for determining the manner of data processing, is responsible for establishing at least one of six possible legal bases for processing personal data including consent, performance of a contract, legitimate interest, legal requirement, and public interest. I will review the three most relevant legal bases that image diffusion models such as Stability AI and Midjourney could use to justify personal data collection.

4.2.1 Consent

Both the GDPR and CCPA offer consent as a legal basis. However, many open-source models, including Stability AI, have faced criticism for web scraping data without obtaining the necessary consent from individuals, allowing users to opt out rather than seeking users to opt-in. The appeal of open sourcing a dataset is perhaps contradictory to implementing a process involving obtaining consent. Open source datasets are appealing because of their size (LAION 5B is 5 billion images), their range, and the lack of regulations (Sahu 2022). Obtaining specific consent, especially for images already circulated, can be a costly and time intensive process. In addition, the size of such a dataset considering that at least some data subjects will not give their consent will diminish, which creates the added negative effect of reducing the number of images the model can learn from, making it a worse model.

Though there is an incentive for computer scientists to disregard consent, a system in which there is no option for artists to deny consent poses the risk of disincentivizing sharing art online. Artists such as Steven Zapata have indicated that some artists intend to cease posting their work online in an effort to not have generative image models learn their style. Because these models are only as good as their training data, it is in companies' long-term interest to foster an environment where artists feel confident that they can share their art online while preserving their right to deny access to their art for the purpose of downstream model training.

Although it is important to gather consent, privacy scholars have acknowledged how our present implementation of notice and consent has its faults. As the World Economic Forum has noted, the prevalence of consent policies, the inaccessibility of privacy policy language, the length of notices, and the inability to receive the same service while denying consent has effectively "placed the burden of privacy protection on the individual"(Ly, no date). Furthermore, the design of notice and consent assumes ideal conditions (unlimited time, sufficient knowledge, and a lack of service dependence) when making privacy decisions (Kröger et al. 2021).

The status quo of requiring artists to go out of their way to opt out undeniably places this burden of maintaining standards of privacy protection on data subjects rather than the data processors. In the context of generative image models, the potential harm of extracting data consensually falls on the data subject (Viljoen 2020); artists can have a diffusion model learn their style before they are able to develop it themselves, or models can replace artists in their own market. If we were to prioritize harm reduction in

personal data collection for image diffusion models, using the consent based approach only works if data processors make notice and consent policies accessible, short, and without manipulative designs. In the future, groups seeking to web scrape with consent in mind should assume that consent is not given as the default and that imperfect conditions exist.

4.2.2 Contract

Another relevant legal basis could be the contractual legal basis, which is when "the processing is necessary for a contract you have with the individual, or because they have asked you to take specific steps before entering into a contract" ("Lawful basis for processing" 2022). Adtech companies, such as Meta, started off by attempting to use contract as their legal basis by claiming that collecting personal information is necessary to deliver personalized ads. The Ireland DPA showed that they intend to hold companies to a higher standard of honesty in what they claim their contractual necessity to be; Meta was denied the ability to use this legal basis because behavioral advertising is not its foremost purpose.

This legal basis could be used by generative image model companies should they decide to switch from web scraping to curating data sets directly from artists and photographers. For example, if the model's developer has entered into a contract with photographers to use their images as part of the training set, the developer may use the contractual basis as the legal basis for collecting and processing the images.

4.2.3 Legitimate interest

One of the most commonly used legal bases used to justify personal data collection is legitimate interest. This legal basis is the most flexible of the legal bases, as it can be used for marketing, fraud prevention, or network security. However, its flexibility also means that such a legal basis has the largest burden of proof; you must provide justification why there is (1.) a legitimate interest in processing the data subjects' data, (2) the processing of data is necessary to perform such interest, and (3) the legitimate interest is worth sacrificing the data subjects' rights ("Lawful basis for processing" 2022).

Recent cases have indicated that courts in the European Union are placing a higher emphasis on stipulations 2 and 3 that models such as Stability AI and DALL-E may be unable to prove should they choose to use legitimate interest as their basis. Open AI has claimed its data collection falls under the basis of legitimate interest, though data protection officers are skeptical in how their companies' interests measure up to the sacrificing of rights (Heikkila 2023). Italy famously banned ChatGPT for the lack of a legal basis for data collection as well as an inability to explain how personal data was being used post-collection (McCallum 2023). Generative image model developers will have a difficult time making the claim that they have a substantial legitimate interest that is worth sacrificing the artists' rights and choice being taken from them, especially as European authorities are moving towards a consent based approach that emphasizes the ability of citizens to choose (Heikkila 2023).

4.3 Personal data, legal rights

Both regulatory frameworks give data subjects certain rights to their personal data throughout the processing system. The GDPR and CCPA generally have similar legal rights granted to data subjects/consumers protected under the law, but the GDPR extends several legal rights that are not included in the CCPA (Jehl et al. 2018). The most relevant legal rights for this discussion that both provisions guarantee are the right of access and the right to erasure.

4.3.1 Right to erasure

Most blatantly irreconcilable with generative image models is the right to deletion/the right to be forgotten, which gives individuals the right to request the deletion or removal of their personal data from the systems of data controllers. However, complying with this right can be particularly challenging for machine learning developers. This is because models are often trained on large datasets of images, which can include a vast number of images containing personal data. In order to comply with the right to deletion, developers would need to identify and delete all instances of an individual's personal data from their entire dataset. However, this may be a difficult and time consuming process for images whose text-image pair does not include the artist's name. Fortunately, the easier it is for an image to be identified with an individual, the easier it is to identify said image and remove it from the dataset. As a result, artists with the most personal data will also have the strongest ability to exercise their rights to erasure.

Though companies could simply remove said images from their dataset and retrain their model, this suggestion does not take into account cost considerations, as computation is the most cost-intensive part of the process. Though having a model unlearn certain images is theoretically possible in very particular cases, de-learning is currently an open problem in its infancy of being understood. However, new methods around AI model disgorgement, or the removal

of influence of specific training data, have been published and are in its early development. Particularly, differential privacy and dataset emulation are two possible remedies that "proactively [train] a model in a way that provides a mathematical proof that no particular piece of training data had more than a negligible effect on the model or the content it generates" and "captur[ing] respectively the general distributional properties of the original training set while maximizing the geometric, perceptual, or conceptual distance from [the training set]" (Achille et al. 2023), respectively. While these new methods offer some hope in providing an easier pathway to reconciling these models with the right to erasure, these methods require more research before policy can mandate implementation.

Fulfilling the right to erasure or the right to opt out (the two are being grouped together because the effect of adhering to such legal rights means not including an individual's image from the dataset, in actuality or in consequence) requires transparency regarding who consumers can ask to honor these rights. Stability AI points to LAION when individuals have asked to opt out/deleted, and the confusion has meant individuals are unsure of where to go for remedy (Xiang 2022). Though LAION 5B is the dataset that indexed their images, it is the Stable Diffusion model that artists may want their work to not be trained on, and this nuance is lost on both of the companies, allowing neither to take responsibility.

4.3.2 Right to access

Generative image model developers may also find difficulty in complying with the right to access, which requires that data processors provide data subjects with the details of what personal information has been collected. The barriers a generative image model developer would face in complying with the right to access is similar to that in the first step of the right to erasure, identifying all of the personal information of the individual, though this time the purpose is to relay to the consumer what images have been collected.

5 The monopolization problem

Whether we allow for open sourcing of images depends on the classification of such data in the aforementioned public/personal categories. If regulatory agencies do side with companies like Stable Diffusion, then we can examine the competition and privacy tradeoffs for this decision. The public/private classification of data as specified by GDPR and CCPA has a significant downstream effect on the competitive landscape and the applicability of privacy regulations. Additionally, the current conception of such a distinction is dependent on decisions that regulatory agencies make as technologies push the boundaries of the meaning of regulations. Thus, it only makes sense that such a decision is not made in a vacuum of outdated black letter criteria, but also within the broader context of privacy norms, expectations, and consequences.

5.1 Open sourcing images

The argument for maintaining the status quo method of obtaining images for datasets is as follows: open-sourcing existing images encourages competition by lowering barriers of entry for new companies seeking to create generative image models. The tradeoff is a classic one that the technology industry constantly faces: increased competition at the cost of privacy.

As Microsoft invests $10 billion into Open AI (Bass 2023) and image generator startup Imagen receives $135 million in funding with Google as an investor (Imagen 2022), the rapidly growing generative AI industry will likely remain dominated by tech companies with deep pockets and existing infrastructure. These companies enjoy the benefits of scale–they have the pay and namesake to draw such high level talent and the organizational capacity to direct a group in producing complex algorithms.

Despite the potential benefits to having industry leaders dominate the AI arms race, closing a profitable market from competitors remains a dangerous prospect due to the potential of future abuses of power and a reduction in customer choice. This exemplifies a claim made by law professor Tejas Narechania: models are naturally monopolistic because of the cost to create and optimize them. Adding the extra barrier of monetizing an image dataset by disallowing open sourcing images only further prevents small startups with a potential to disrupt the market from entering. Open sourcing datasets has been hailed amongst both industry leaders as well as champions of antitrusts: industry leaders enjoy its ability to cut costs and the model improvements from diversified, expansive training sets; and antitrust proponents appreciate lower barriers to entry.

Open-sourcing is not the only way to prevent monopolies from emerging in the technology space. Antitrust advocates typically look to apply long-standing antitrust regulations to naturally monopolistic tech phenomenons. In his same piece, Narechania went on to explain: "if these applications do indeed act as natural monopolists – or even, perhaps, as mere monopolists or oligopolists – then … the legal traditions of natural monopoly regulation, and market

regulation more generally, offer a ready, but perhaps overlooked, framework to help address these problems" (Narechania 2021).

The implications of assessing this data as public poses significant privacy concerns. Privacy protections will be narrowed towards an antiquated understanding of identifiability that fails to consider the evolving ease of identifying an individual. Furthermore, the classification of this data as public fails to account for the expectations an individual may have regarding their privacy when uploading images of their art and the demands from artists that their consent be considered.

Accepting an open source model would necessitate a public classification of training data, implying that companies are not required to adhere to privacy procedures such as obtaining a legal basis for data collection or granting data subjects rights over their data. In spite of no legal requirement to honor the right of erasure, companies like Stability AI and Midjourney have claimed they will implement an opt out function for their next generative image model. However, until further legislative or judicial action clarifies the legal requirements for companies maintaining large datasets, there will continue to be ambiguity surrounding opt out, and companies will remain disincentivized to gather meaningful consent.

5.2 Licensed or curated dataset

Encouraging a system that necessitates licensed or privatized datasets runs the risk of more deeply entrenching the market position of the market players already in dominance. If the legality of input datasets hinged on whether the model trained on them had acquired a license, the cost of curating such datasets may result in wealthy market players like Microsoft or Google being some of the only companies that are able to invest the time, money, and legal resources into doing so. Microsoft, Google, and other companies seeking to create a licensable dataset would have had to obtain consent or adhere to some other legal basis in order to have collected the images for their dataset. Companies seeking to train generative image models would then have to pay these companies who had licensed their datasets. Privatizing datasets may provide privacy benefits by creating a market for images in these datasets. Money raised from this privatization could be used to repay artists for opting into being included into a personal dataset.

Privacy implications may run the other way as well. By decreasing transparency, companies could web scrape without the oversight of the public. Legal challenges concerning fair image use were possible because of the open nature of the datasets and the ability for the public to review how the images were sourced. Finally, offloading privacy to the data collectors may still mean that the same issues arise when obtaining consent or establishing legitimate interest as the data collectors' legal basis.

6 Conclusion

It's important to note that the alternatives in this piece are not exhaustive nor are they necessarily prescriptive. At this stage, it is difficult to imagine a viable alternative given the complexity. However, it's crucial that the lack of an easily understandable solution does not discourage examination of privacy frameworks as new technologies that do not fit squarely within our frameworks arise.

As image generative models circumvent copyright laws through the fair use clause, I look at how privacy laws could be applied if artists' works are imagined as personal, rather than public data. Legal bases would be required for such a collection. This could provide artists a protection against having their images used against their wishes, especially if companies choose to use consent as their legal basis. Additionally, the requirement of providing individuals legal rights to their imaging data could mean a renewed dedication to tackling the delearning problem or an increased dedication to ensuring consent was given in the first place so as to mitigate the amount who opt out later.

Accounting for privacy matters does come at a technical cost; inhibiting image datasets from being interpreted as publicly available will likely decrease its size and range, hurting the training capability of these models. However, it should be noted that the quality of artwork artists create already determines the quality of the model. Accounting for privacy concerns of the artists whose works determine the quality of the model may come at a technical cost, but perhaps a necessary one, given the extent that artists are stakeholders.

Lastly, interpreting these datasets as public or not creates monopolization implications. Though privatizing these datasets may appear to further entrench the market position of dominant players and restrict competition, remedies through antitrust law and other state regulation for such instances can be implemented as a preventive measure.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: The author has accepted responsibility for the entire content of this manuscript and approved its submission.

Use of Large Language Models, AI and Machine Learning Tools: None.
Conflict of interest: The author states no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

References

Achille, A., Kearns, M., Klingenberg, C., and Soatto, S. (2023). AI model disgorgement: methods and choices. *Proc. Natl. Acad. Sci. U.S.A.* 121: 4–6.

Art. 4 GDPR – Definitions (2016). General data protection regulation (GDPR), Available at: https://gdpr-info.eu/art-4-gdpr/ (Accessed 8 May 2023).

Baio, A. (2022). Exploring 12 million of the 2.3 billion images used to train stable diffusion's image generator, Available at: https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/ (Accessed 30 August 2023).

California consumer privacy act: a practice overview (2020). American Bar Association, Available at: https://www.americanbar.org/groups/litigation/committees/corporate-counsel/practice/2020/california-consumer-privacy-act-a-practice-overview/(Accessed 8 May 2023).

FTC issues opinion and order against Cambridge Analytica for deceiving consumers about the collection of facebook data, compliance with EU-U.S. Privacy Shield (2019), Available at: https://www.ftc.gov/news-events/news/press-releases/2019/12/ftc-issues-opinion-order-against-cambridge-analytica-deceiving-consumers-about-collection-facebook (Accessed 5 December 2019).

Heikkila, M. (2023). OpenAI's hunger for data is coming back to bite it. *MIT Technology Review*.

Hill, K. (2023). This tool could protect artists from A.I.-Generated art that steals their style, Available at: https://www.nytimes.com/2023/02/13/technology/ai-art-generator-lensa-stable-diffusion.html (Accessed 13 February 2023).

HiQ Labs v. LinkedIn (2022). Available at: https://en.wikipedia.org/w/index.php?title=HiQ_Labs_v._LinkedIn&oldid=1128822683 (Accessed 2022).

Hill, K. (2023). This tool could protect artists from A.I.-generated art that steals their style. *The New York Times*, Available at: http://www.nytimes.com/2023/02/13/technology/ai-art-generator-lensa-stable-diffusion.html (Accessed 13 February 2023).

Imagen technologies—funding, financials, valuation & investors (2022). Available at: https://www.crunchbase.com/organization/imagentechnologies/company_financials (Accessed 7 May 2023).

Jehl, L., Friel, A. and Llp, B. (2018). CCPA and GDPR comparison chart.

Kröger, J.L., Lutz, O.H.-M., and Ullrich, S. (2021). The myth of individual control: mapping the limitations of privacy self-management, Available at: https://doi.org/10.2139/ssrn.3881776.

Lawful basis for processing (2022), Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/ (Accessed 17 October 2022).

Ly, J.U. (no date). Reimagining notice & consent for human-technology interaction.

McCallum (2023). ChatGPT banned in Italy over privacy concerns. *BBC*, https://www.bbc.com/news/technology-65139406.

Microsoft Invests $10 Billion in ChatGPT maker OpenAI (2023), Available at: https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai (Accessed 23 January 2023).

Narayanan, A., Pasov, H., Gong, N.Z., Bethencourt, J., Stefanov, E., Richard Sin, E.C. (2012). On the feasibility of internet-scale author identification, Available at: https://doi.org/10.1109/SP.2012.46.

Narechania, T.N. (2021). Machine learning as natural monopoly, Available at: https://doi.org/10.2139/ssrn.3810366.

Purtova, N. (2018). The law of everything. Broad concept of personal data and future of EU data protection law, Available at: https://doi.org/10.1080/17579961.2018.1452176.

Sahu, A. (2022). Why open-source companies have the competitive advantage, Available at: https://www.weforum.org/agenda/2022/08/open-source-companies-competitive-advantage-free-product-code/ (Accessed 17 August 2022).

Tene, O. and Polonetsky, J. (2013). Big data for all: privacy and user control in the age of analytics. *Northwest. J. Technol. Intellect. Prop.* 11: 239.

Viljoen (2020). A relational theory of data governance. *The Yale Law Journal* 131: 19–22.

Xiang, C. (2022). AI is probably using your images and it's not easy to opt out, Available at: https://www.vice.com/en/article/88q9gp/ai-is-probably-using-your-images-and-its-not-easy-to-opt-out (Accessed 26 September 2022).