
CARROLL AND
MILTON PETRIE
EUROPEAN
SCULPTURE
COURT

PEER-REVIEWED ARTICLE

MEMORY INSTITUTIONS MEET AI: LESSONS FROM CRITICAL TECHNOLOGY DISCOURSE

JORDAN FAMULARO AND REMI DENTON

ABSTRACT | Galleries, libraries, archives, and museums (GLAMs) across the globe are building new datasets to render their collections open, machine-readable, and internet-accessible. The new generation of GLAM datasets have wide reach, offering to turn institutions inside-out so that remote audiences can view, download, share, and remix digital assets. GLAM institutions have treated the associated turn to open data as inherently positive—able to promote cultural understanding and appreciation in ways that promise scale, accessibility, and customization. However, some critics suggest that upsides of technology for GLAM datasets need to be balanced with risks that can arise from their design, development, and integration into artificial intelligence (AI) technologies. In this work we ask: how should GLAMs account for the emergence of AI-driven experiences built upon GLAM datasets? We seek to answer this question by flagging key ethics and governance issues in tandem with supplying some guardrails for navigating them. We examine GLAM datasets from a sociotechnical perspective. Drawing on our experiences as researchers spanning multiple areas (computer science, computer vision, AI ethics, art history, and cybersecurity) and working in different sectors (industry and academia), we identify salient concerns and remediations from critical technology discourse on dataset development for AI systems.

KEYWORDS | GLAM institutions, artificial intelligence, collections, data, accessibility

Introduction

Galleries, libraries, archives, and museums (GLAMs) across the globe are building new datasets to render their collections open, machine-readable, and internet-accessible.[1] GLAM institutions have largely treated the turn to open data as inherently positive, able to promote cultural understanding and appreciation in ways that promise scale, accessibility, and customization.[2] Indeed, GLAM datasets have wide reach, offering to turn their institutions inside-out so that collections may be remotely viewed, downloaded, and shared by anyone with an internet connection. In the age of artificial intelligence (AI), GLAM datasets can enable new forms of audience engagement with collections.[3] AI creates possibilities for machines to present recommendations, provide predictions, generate content, and display curated information to remote audiences—with more conceivable functions to come as the technology evolves. However, some researchers and artists suggest that the upsides of GLAM datasets need to be balanced with risks that can arise from their design, development, and integration into AI technologies.[4] Whereas many GLAMs have marshaled the technical expertise and resources required to remodel their collections into digital datasets, in this paper we seek to spark public discussion about the conditions and guardrails that should guide the evolution of these projects in the age of AI.

In this work we ask: how should GLAMs account for the emergence of AI-driven experiences built upon GLAM datasets? We seek to answer this question by flagging key ethics and governance issues in tandem with supplying some guardrails for navigating them.

Our approach focuses on GLAM datasets as sociotechnical systems. Drawing on our experiences as researchers spanning multiple areas (computer science, computer vision, AI ethics, art history, and cybersecurity) and working in different sectors (industry and academia), we identify salient questions from critical technology discourse on dataset development for AI

systems. Examining GLAM datasets from a sociotechnical perspective foregrounds the need for cultural memory institutions not only to make their digital systems technically workable, but also to confront the implications of cultivating new participatory communities through datasets that can be viewed, downloaded, manipulated, and reused through the internet.[5] Further, institutions must shape new accountability assurances when they allow digital collections to be distributed and recontextualized beyond their own website by users working in a wide matrix of machine-to-machine linking and processing of data.[6] While generative AI systems such as ChatGPT, DALL-E, and Stable Diffusion gain public fascination, scholars are investigating the landscape of risks and potential guardrails to guide the responsible development of these technologies. Yet the AI systems' reliance on data from GLAM datasets remains largely overlooked.

First, we align our core analysis with common stages of dataset development in order to highlight the many decisions involved from start to end. Next, we embed the core analysis into a wider institutional context by examining how technologists conceptualize, make, monitor, use, and secure datasets. Tactically, this move emphasizes that datasets are made and perpetuated through a series of choices and actions by GLAM staff and their consultants, partners, and audiences. The making and stewarding of GLAM datasets thus becomes the central concept shaping how we organize our exposition and recommendations. Our evidence consists of primary sources (internet sites on the public web; e-mail communication with The Met; mainstream journalism) and secondary sources (a literature which we call critical technology discourse, combining scholar and practitioner perspectives, integrated by us with perspectives from digital art history and museum, library, and archival studies).

We present five recommendations to guide ideation and development of GLAM datasets in line with advocacy for responsible dataset development for AI in research communities. Our discussion reveals some weaknesses in datasets already produced by GLAMs in different corners of the world, in addition to some concerns about their use. To throw the issues into relief, we describe them with reference to one prominent example, The Metropolitan Museum of Art's Open Access initiative.[7] The museum configured its Open Access dataset to share information about hundreds of thousands of artifacts for unrestricted commercial and noncommercial use beginning in 2017.[8] The effort relied on techniques from machine learning (ML, an approach used in AI science) and a subset of ML known as computer vision.[9] The Met's dataset is an impressive early accomplishment for the field, yet it also provides cautionary lessons that need close study from GLAM professionals, boards of directors, and scholars in the humanities, social sciences, and computer science.[10]

Although AI technology may be unfamiliar terrain for some readers of this journal, there is shared ground with respect to how computer scientists and GLAM professionals work with collections that need to be designed, organized, interpreted, and made accessible for others. Our recommendations are intended to invite conversation, not to be the final word.

Where GLAM Institutions Meet AI

One can trace a history of AI-compatible cultural memory collections to open information access initiatives by a number of prominent GLAMs and, more broadly, collections-as-data projects. A central tenet of "open" GLAMs—that digital reproductions of public domain works should be made available for anyone to freely access, use, modify, or share[11]—has given rise to large internet-based collections that can form the basis of AI datasets. According to a 2022 Open GLAM Survey, more than 1,400 GLAM institutions worldwide have released digital collections on open access terms.[12] These include some of the 3,700+ institutions that participate in the European Union's Europeana,[13] the J. Paul Getty Museum[14] (Los Angeles), National Palace Museum[15] (Taipei), Cleveland Museum of Art,[16] Art Institute of Chicago,[17] Smithsonian Institution[18] (Washington), National Gallery of Art[19] (Washington), Powerhouse Collection of the Museum of Applied Arts and Sciences[20] (New South Wales), and The Met[21] (New York). Treating collections as data in cultural memory institutions is an adjacent and in some ways complementary movement to encourage computational use of digitized and born-digital collections.[22] But our examination of AI-related initiatives at GLAMs across the globe in recent years suggests that more is at stake than simply democratizing access and fostering new uses of collections.

Development of AI systems using GLAM datasets has involved a constellation of external groups operating with different incentives. External GLAM relationships include partnerships with large corporations, such as Google[23] and Microsoft,[24] data hosting by tech companies, such as AWS (Amazon Web Services),[25] database integration with businesses, such as Artsy[26] and nonprofits, such as

Wikimedia Foundation,[27] software development platform services by commercial firm GitHub,[28] collaborations with universities and schools, such as Parsons School of Design,[29] analytics arrangements with data science companies,[30] data aggregation services by NGOs and public organizations,[31] and crowdsourced labor by dispersed and sometimes anonymous contributors.[32] This list of organizational ties suggests that a myriad of imperatives come together to shape GLAM datasets and AI projects built from GLAM datasets. Stakeholders lay claim to different aspects of a dataset project, possibly even the data itself.

The Met provides an example of how AI projects made with GLAM datasets give rise to a broad partnership ecosystem with for-profit and nonprofit organizations. The Met's dataset has anchored collaborations with Microsoft,[33] Massachusetts Institute of Technology,[34] Parsons School of Design,[35] Wikimedia Foundation, University of Virginia School of Data Science, Google, Pinterest, and Creative Commons.[36]

GLAM datasets, in addition to being reliant on multiple organizations, have public-facing roles and effects. For example, the Met has made its dataset available to the public in two ways, through an application programming interface (API)[37] and a comma-separated values (CSV) file available on Github.[38] While the API enables the museum to open up its data and functionality to third parties by means of an intermediary layer between a web server and application, the CSV file allows data scientists, artists, and others to integrate the data easily into external datasets and technologies[39] (subject to The Met's published terms and conditions).[40]

In the present AI age, the new openness of GLAM datasets like The Met's compels cultural memory institutions to confront the spread and reuse of collection data in a wide matrix of internet activity beyond their strict control. This is particularly salient given the emergence of new generative image technologies, such as Stable Diffusion and DALL-E, that are made with data scraped from the internet. Some of these AIs are fueled by publically available GLAM datasets: for example, Stable Diffusion was trained on data from The Met's dataset.[41] There are few governance systems in place for generative AI systems, which carry risks that have had little time to be considered while firms compete in an arms race to deploy. As told by the CEO of OpenAI, the company behind DALL-E: "The current worries that I have are that there are going to be disinformation problems or economic shocks, or something else at a level far beyond anything we're prepared for."[42]

Digital heritage is no longer a practice of making technical systems work but also a practice of facing new sociotechnical outcomes, as Ross Parry claims, such as recontextualization of GLAM collections by other actors and by machine-to-machine processing of data.[43] As such, we suggest that the horizon of social responsibility widens as GLAMs address how their datasets will integrate with AI systems made by developers either inside or outside the institution. As just one example of what GLAM governance may now be expected to address, an explainer video for DALL-E 2 published by its commercial developer, OpenAI, features an image of Leonardo da Vinci's *Mona Lisa* with a mohawk, showing how the AI allows a user to select an existing image and give the command, "give her a mohawk" (https://openai.com/dall-e-2). It takes little stretch of the imagination to envision how internet users may use GLAM datasets and generative AI systems like DALL-E to produce images that are abusive, harmful, or extremist while retaining some visual semblance of the underlying works in GLAM collections.

In the remainder of this article we explore potential guardrails that could guide the evolution of GLAM datasets in the age of AI, drawing upon prior scholarship advocating for higher ethical standards of AI dataset development and stewardship. Datasets, more broadly, have been front and center of the controversies surrounding AI, being implicated in debates about bias, fairness, consent, and agency with respect to impacted stakeholders.[44] Some of the contexts giving rise to these controversies are readily recognized as posing major risks to health and safety (such as autonomous vehicles). Of more concern to GLAMs and their publics, we suggest, are threats to social cohesion and fundamental rights, which are associated with AI datasets because of how they can condition behavior and thought, limit individual autonomy, and worsen inequality. Indeed these dynamics within global digitalization and algorithmization have been understood as systemic risks.[45] For example, biased, unfair, and discriminatory AI has been partially attributed to an over-representation of white, western, male perspectives in AI datasets;[46] controversies regarding the use of publically available text and image data scraped from the web have sparked new debates regarding attribution, consent, and data subject agency;[47] and the regular reliance on crowdsourced workers within dataset development pipelines has raised a myriad of ethical concerns regarding labor conditions.[48] In response, a polycentric community of scholars and practitioners has stepped forward to provide guidelines and tools in the form of soft governance, and diverse researchers have established a rich body of

scholarship around critical issues pertaining to AI datasets.[49] These reparative developments, which we refer to as critical technology discourse, offer valuable insights for the GLAM sector that we distill into our recommendations below.

Making GLAM Datasets Responsibly

To make a dataset is to engage in ethical choices that impact a broad set of stakeholders. GLAM datasets take shape from a combination of contributions from inside and outside the institution. Mass digitization projects at cultural memory institutions, while formally seeking to serve the public interest, are infused with diverse and even conflicting political and economic motives.[50] AI systems derived from GLAM datasets subsequently inherit and embed new, value-laden norms, assumptions, and design decisions that reflect power dynamics and human labor underlying their production.[51] GLAM datasets reflect a series of design choices and power structures that have built the institution's physical collection over time, extending this history into the present.[52] In turn, the datasets shape an expanse of experiences for internet "prosumers" who not only consume content but produce new engagements by downloading, uploading, sharing, tagging, and remixing.[53] In the age of AI, how GLAMs will develop new policies and practices to deal with these chains of stakeholders and their interactions is a rich question that merits fuller discussion.

GLAMs' policies and procedures have traditionally been opaque to outsiders, but societal pressures might force this to change. Twenty-first century ethics in the GLAM sector is built upon new theory and practice of transparency, a self-reflexive mode of communication that admits accountability and discloses the stakes of decision-making.[54] We propose that AI-driven experiences built on GLAM datasets will accelerate the shift. This is because the relationship between dataset development and outcomes mediated by AI presents novel uncertainties and poorly understood risks.

GLAMs need to contend with two facts: the concept of risk captures a growing role in how society establishes guardrails for digitization of contemporary life, and institutions face new pressures to govern technology risks.[55] Sociologists Ulrich Beck and Anthony Giddens theorized that contemporary society is a "risk society," a term concerned with the transition from industrial society to the current era shaped much more by technological hazards.[56] The risk society is distinguished not only by distribution of "goods" (wealth) but more so by distribution of "bads" (technological hazards produced by society such as misinformation, online abuse, and cyberattacks). AI risk management is now emerging as an important field of research and strategy across sectors.[57]

We suggest that GLAMs' contribution to the development of AI systems through the datasets they develop and make available on the internet will intensify pressures on GLAMs to be transparent about choices and actions in dataset development. This is partly because risk-based governance—which is marked by transparency-centered regulation mechanisms, such as audit, reporting, and risk assessment—has come into favor in regulation of digital society and digital markets.[58] While this is the case in "hard" governance, i.e., legally binding regulation, there are parallels in "soft" social norms that steer conduct, as we discuss further below. As GLAMs continue to integrate their datasets with digital society, they should rethink their self-governance policies and procedures with particular attention to documentation and disclosure.

GLAMs, as institutions that serve the public interest, ought to be norm-builders in regard to evaluating the potential impact of datasets they create. Although this territory is new for GLAMs, industry expectations are taking shape around the notion that, before and during the development of any new service or project involving information technology, organizations seeking to lead institutional responsibility norms should perform an ethical impact assessment[59] or similar evaluation to examine societal consequences of the design. Similar to an environment impact assessment that precedes a real estate development project, an impact evaluation framework for GLAMs would enable more informed choices about ethical and social implications, although research and consensus-building are needed to establish an accepted procedure for memory institutions.

Recommendation 1: Conduct an impact assessment about interaction between dataset decision choices and knowledge paradigms or biases.

While impact assessments are important exercises in their own right, organizations typically conduct them internally and need to take extra steps to provide means of external accountability. Documentation that communicates results of the impact assessment and other important decision points in dataset development is, according to a growing technology ethics literature, owed to stakeholders outside

Figure 1. *Jain Svetambara Tirthankara in Meditation*, ca. 1000–50. Marble, 99 cm height. The Metropolitan Museum of Art, New York, Purchase, Florence and Herbert Irving Gift, 1992, www.metmuseum.org. Creative Commons CC0 1.0 Universal Public Domain Dedication.

the organization.[60] Collections-as-data approaches have proposed standardization of documentation, such as data packets that bundle transcriptions with contextualizing information to explain decisions made while creating data.[61] In AI-specific literature, there is increasing awareness that rigorous and transparent dataset documentation can help mitigate ethical concerns in AI dataset development and use.[62] For example, documentation serves to hold dataset developers accountable for their decisions, enable dataset users inside and outside GLAMs to make responsible decisions regarding safe and appropriate use, and allow third-party researchers to offer validation or critique as part of scholarly practice. A standardized dataset documentation framework, tailored to the GLAM sector, would support and complement an impact assessment by guiding design decisions and making them visible to a broad set of stakeholders.

Recommendation 2: Produce comprehensive documentation, including decision provenance, of how the dataset is made. Make the documentation available for review outside the institution.

There are models on which the GLAM sector can draw. Within AI research and practice, there is growing awareness of the importance of public transparency and there are extensive efforts to encourage public-facing documentation of models[63] and datasets.[64] Researchers have recognized the gap in guidance for documentation of arts-related dataset development. For example, Artsheets is a research-based and practice-oriented framework that offers a checklist and questionnaire to guide arts-based dataset documentation efforts with specific focus on ethical, social, cultural, legal, and historical considerations.[65] We do not fetishize transparency,[66] but we do adopt the working premise that more publicly available information is better than less, all else being equal. Moreover, we advocate for transparency as an important approach to enabling multi-stakeholder engagement with GLAM datasets and technologies.

We requested a documentation file from The Met in order to gain the chance to evaluate dataset design choices. The response directed us not to what we requested but to alternatives: the Terms and Conditions governing the dataset's use, and articles and FAQs on the museum's website pertaining to the Open Access initiative.[67] This decentralized array of information makes it difficult for researchers and the public to understand how the dataset was made and by whom.

To improve transparency about GLAM decision-making and to guide implementation of our first two recommendations, the following discussion distills three themes that merit exploration toward the formation of an impact assessment and documentation framework for the GLAM context: data classification, cleaning, and annotation.

Data Classification

Classification systems serve an important purpose; they lend order to the incredible complexity of the world. These systems are an essential component of turning GLAM collections into machine-readable datasets. Yet, as Geoffrey Bowker and Susan Leigh Star remind us,[68] classifications embed politics. When making GLAM datasets, choices of which categories to include and how to sort data are critical design decisions that shape the final dataset by making some aspects of the underlying collection legible and other aspects illegible or lost entirely. Furthermore, classification decisions compound over time, since varying labels can be added at different moments and by different people. Now that citizen scientists and crowdworkers contribute to digital tagging efforts—supplementing the contributions of GLAM professionals—classification work is done both outside and inside the institution, as we discuss further below. In turn, the aggregate of classification decisions creates affordances for certain types of research, learning, and experience, while excluding or inhibiting others.

The Met dataset contained 54 columns (or divisions) of information when examined as a CSV file at the time of our investigation. A significant portion—11 columns out of 54—conveyed information for each museum artifact about the biography of the artist.[69] At the same time, there were zero columns reserved for other important aspects of an object's history and cultural meaning, such as its state of preservation. The proportion 11/54 represents an inbuilt intelligibility that invites users to explore more questions about the artist of each object and fewer about its ritual status, affective presence, physical state over time, and a number of other issues that remain suppressed because of the choice about which classifications are available in the CSV file.[70]

Classification also strips away complexity in favor of neat categories. One 11th-century sculpture at The Met (Figure 1), among many other examples, illuminates the erasure of context when a classifying process seeks to render "title" into a compressed unit. The museum's traditional title for the sculpture is *Jain Svetambara Tirthankara in Meditation*, but its title in the digital dataset is *Figure*. Far from a rare occurrence, it represents a broader design strategy wherein comparable sculptures in The Met's Asian Art Department (and other departments) have undergone a reduction from a title that is long-form in the humanities-based domain to the single word *Figure* in the CSV file.

Consider how the computational title *Figure* erases interpretive context. The sculpture's traditional title evokes the object's religious derivation (Jain), color (*svetambara* or white-clad), devotional association (*tirthankara* or ford-crosser), and attitude (in meditation).[71] It is unsurprising to find that other scholarly authors employ slightly different terms for the same object, such as *Seated Jain Tirthankara*,[72] reflecting different interpretive and expressive priorities. This example suggests that, whereas humanities methods allow categories like "title" to remain contingent and thickly descriptive, computational methods freeze them into relatively barren units, often a shortened form. A potential epistemic effect of data classification, therefore, is foreclosure of debate about points of interpretation that may be conditional or provisional.

Our intention here is not to suggest that there is a perfect set of column headings for The Met's dataset; rather, the dataset makers bear some responsibility for explaining and defending their choices. Consultation with domain experts and stakeholder groups would be a prudent step at the design stage in order to explore how classifications affect both user experience and community representation by restricting inquiry and understanding in some ways while expanding them in other ways. In light of these considerations, we recommend:

Recommendation 2a: In the impact assessment and public-facing documentation, include information about consultation with domain experts and stakeholder groups to explore how classifications affect user experience and community representation.

Data Cleaning

Broadly speaking, data cleaning refers to a set of activities designed to improve the quality of data and convert the results into a form that is appropriate for modeling and interpretation. While precise data cleaning processes tend to vary across context, common activities include removing outliers, resolving redundancies or duplications, and correcting or removing mislabeled data. Data cleaning often embeds assumptions

regarding what constitutes "high quality" data and subjective decisions regarding which attributes and variables will be counted versus ignored.[73] For example, consider a dataset where individual images have been classified or labeled by aggregating multiple judgments from different people (e.g. judgments about artistic style of a painting). In cases where there are highly divergent opinions, a dataset developer must make a choice regarding how to proceed, e.g. removing entirely, relabeling, or giving an existing label greater weight. Each choice carries value-laden implications for the final dataset.

In short, data cleaning reflects layered power relations, organizational restraints, sociotechnical norms, individual habits, technical infrastructure, and happenstance. In light of these considerations, and motivated by existing calls to document and communicate data cleaning processes,[74] we recommend:

Recommendation 2b: In the impact assessment and public-facing documentation, record assumptions and procedures for data cleaning.

Crowdsourced Annotation

Annotating, often called tagging or labeling, adds extra information to the data and can be performed manually or via rules engines.[75] To find people to perform annotation, dataset developers frequently leverage crowdsourcing platforms, since they cheaply distribute tedious annotations across thousands of gig workers. This method has generated enthusiasm among developers across a number of industry sectors and knowledge fields.[76] Yet there are significant critiques that raise important considerations for GLAMs.

One of the core assumptions underlying the crowdsourcing paradigm is that workers are interchangeable. However, scholars have contested this assumption by studying how the perspectives of individual annotators—shaped by social and cultural identities, familiarity with the problem domain, training, and expertise, and more—influence the resulting dataset labels.[77] There is no industry standard for ensuring that workers have sufficient domain expertise for attaching labels that specialists in the field would consider correct, nor is there wide acceptance of the contingent and relational nature of virtually all data.[78] It is critical for GLAM dataset developers to carefully consider whose perspectives, biases, and values are captured within the annotation process and to document annotation processes for external review.

Crowdsourcing has also been heavily critiqued from a labor perspective. The same conceptual and technical infrastructure that sets up workers as interchangeable also positions them as a faceless, nameless, and largely invisible workforce. Scholars have referred to this form of veiled digital labor as "ghost work,"[79] and, pointing to concerns about exploitation, they have critiqued its compensation and credit attribution practices.[80]

The Met divides its approach to annotations by using one method for the web version of its public digital collection and a second method for the Open Access dataset on GitHub. The longer established of the two is the public digital collection, which exhibits The Met's previous standard of offering annotations made by domain experts only (https://www.metmuseum.org/art/collection). The new approach for the Open Access API and CSV file is to include non-expert tags (https://github.com/metmuseum/openaccess). The museum explained the layperson tagging initiative in 2019 as follows:

With the help of a wonderful human work force, The Met recently added keyword tags to over 275,000 objects in the online Open Access collection. . . . But there are still over 20,000 objects that need to be tagged. As we continue to add more digitized objects to the collection, we need to weigh the cost, time, effort, and accuracy of tagging our objects.[81]

Tradeoffs from annotation come to light in an example from The Met, a 13th-century box called a pyxis (Figure 2). The first challenge has to do with credibility: no public record informs the user whether the tags are human- or machine-generated. The second challenge is the gap in perception that is evident when one compares the pyxis to its tags. In the CSV file, one finds three tags for this artifact: Christ, Man, Donkey. But the total aesthetic effect and historical context of the pyxis are more complicated: the artifact is intercultural, decorated with an amalgam of Christian and Islamic themes. It is a cylindrical box with a lid, a form that derives from ancient Greek ceramic prototypes. Its motifs include Christ, a man, and a donkey because it shows a Gospel scene of Jesus' Entry into Jerusalem—but it also shows a representation of Mary, whose cross-legged posture recalls images of Seljuq rulers, while her turban recalls headgear distinctive to 13th-century Arab communities in Syria and the Jazira. It is brass inlaid with silver, a technique mastered in medieval Syria, where the pyxis is believed to have been made. It is difficult to perceive how the artifact could possibly be signified reliably by the three tags, Christ—Man—Donkey. Rather, the logic of keyword thinking suppresses many aspects of the pyxis.

Figure 2. Unknown artist, *Pyxis Depicting Standing Saints or Ecclesiastics and the Entry into Jerusalem with Christ Riding a Donkey*, ca. 1250–1300. Brass inlaid with silver, 10.5 cm height. The Metropolitan Museum of Art, New York, Rogers Fund, 1971, www.metmuseum.org. Creative Commons CC0 1.0 Universal Public Domain Dedication.

Thus, The Met's statements of support for keyword tags, grounded in a promise of greater accessibility through search, deserve closer scrutiny. For example, The Met's website claims: "The 'depicted subject matter' tagging function gives users a new access point into the collection, allowing them to see how a subject of interest, such as dogs or mirrors, may have been depicted across time, geography, and different types of objects."[82] Missing from this statement are two crucial points: an artwork will remain hidden from search and inquiry if annotated imperfectly, and no artifact will have a perfect annotation.

Moreover, The Met provides no details regarding the annotation process (e.g., guidelines provided to annotators) or the annotators themselves (e.g., recruitment criteria, sociodemographic information, required expertise or training). In addition to raising questions about the particular perspectives embedded in the annotations, The Met's assistance from an unnamed pool of laborers merits further explanation from the museum in relation to ethical questions that have been raised in the literature about compensation, working conditions, and recognition.

In light of the considerations above, here is how we would recommend embedding annotation in the impact assessment and dataset documentation:

Recommendation 2c: If annotation will rely on crowdsourced labor, document why in the impact assessment and external-facing documentation, and create a record of the annotation process, including sociodemographic information about annotators, in order to situate the resulting dataset.

In regard to annotation work for GLAMs, the institution's personnel, board, and audience should demand open dialogue about tradeoffs and defense of the decisions made. One way for GLAMs to facilitate such conversation is to compile a record of all AI annotation projects, then make it available for review by outside parties.

Stewarding GLAM Datasets

Datasets are a novel kind of perpetual collection; they require long-term care from people with skills in information technology, cybersecurity, curation, and pedagogy. GLAMs' stewardship of datasets may combine codes of ethics for memory institution professionals with emerging concepts of stewardship from perspectives of internet ethics, data governance, and consumer ethics. The following discussion emphasizes key decision points for GLAMs drawn from critical technology literature.

Dataset Use: Computation as a Program of Thought

AI and "Big Data" refer not only to the use of tools and processes for developing and analyzing large datasets—they also signify a computational turn in thought.[83] The memory cultivated by the logic of databases and computational methods is far from the type of memory that GLAMs have traditionally committed to foster and preserve. Mike Pepi writes that museums may be "seduced by a transformation into an indexed collection, structured much the way a good database would be: consistent, atomic, scalable, and easily searchable."[84] The shift, which we claim applies beyond museums to GLAMs more generally, leads to a predetermined means of persuasion that Pepi describes as follows: "Information stored in a database is not designed to conjure, remind, or encourage users to look back in retrospect. On the contrary, these data exist to power an application, usually an algorithm that predicts or optimizes future functions."[85]

The ability to analyze data through powerful statistical programs creates many new affordances and opportunities to explore digital data at scale. Yet this mode of engagement with GLAM collections strips away conditions for meaningful interpretation, such as familiarity with cultures represented by texts, art, and artifacts. This important consideration is often lost in the allure around computation of datasets—similar to Alexander Campolo and Kate Crawford's concept of "enchanted determinism"[86]—because of reasons such as terrific speed, efficiency, and rationality.

Now that GLAM datasets are beginning to take hold among institutions with means to build them, it is important to suspend belief in computational enchantments. GLAM professionals need to develop a clear understanding of prevailing norms in computer sciences that produce rationalizing discourses about the capabilities of AI systems.

Recommendation 3: Consider future opportunities and risks presented by the kinds of thinking that computational methods promote, and how such futures might advance or degrade the institution's mission. Foster opportunities and mitigate risks accordingly.

To explain this recommendation, we focus on two consequences of computational thinking: context stripping and seeing patterns where none exist. Both illustrate how datasets are always methodologically joined to human design and bias, as numerous critics of computer and information sciences have argued.[87]

Context Stripping

Computational analysis of real-world phenomena depends on simplification, or modeling, at a distance from the things being studied. Though some commentators refer to the abstraction process as an "aperture"[88] or "flattening,"[89] here we describe it as "context stripping" to emphasize the loss of relevant information. Making data machine-readable means stripping it from its source, necessarily preserving some pieces of meaning and cutting away others by translating complex cultural constructs into simpler forms such as images, text strings, or categorical codes.[90] Dataset development drives context stripping through processes such as classification and cleaning.[91] As a result, the original contexts and communities that produced cultural artifacts become disordered or lost in the data. The stakes of this become particularly salient as AI developers rely on GLAM datasets to fuel emerging generative AI technologies. Critical AI scholars warn of risks of cultural erasure when generative AI simplifies or misrepresents cultural subjects.[92] This particular risk becomes visible in our examination of the CSV file containing The Met's dataset. As discussed above, an 11th-century sculpture (Figure 1) is represented with a simplified title *Figure* in place of the museum's traditional title *Jain Svetambara Tirthankara in Meditation*. This replacement strips important interpretive context from the digital rendition of the sculpture, potentially impacting a user's experience. For example, a user looking for a figure in a meditation pose would not find the digital record of this sculpture by using the CSV file or API. A meditation pose may be more important to some cultures than others, which magnifies the import of the decision to strip away cultural components of the object's description in favor of a brief computational text string.

GLAM professionals should take note of the vast interpretive and experiential gaps that separate physical collections of cultural objects from their digital proxies.[93] Moreover, they should recognize that dataset analysis dictates the shape of these chasms by stripping auxiliary parts of data until they become neat, measurable packages. GLAM dataset development exemplifies how computational thinking is a form of intervention in cultural heritage, which elevates some aspects of cultural memory and suppresses others.

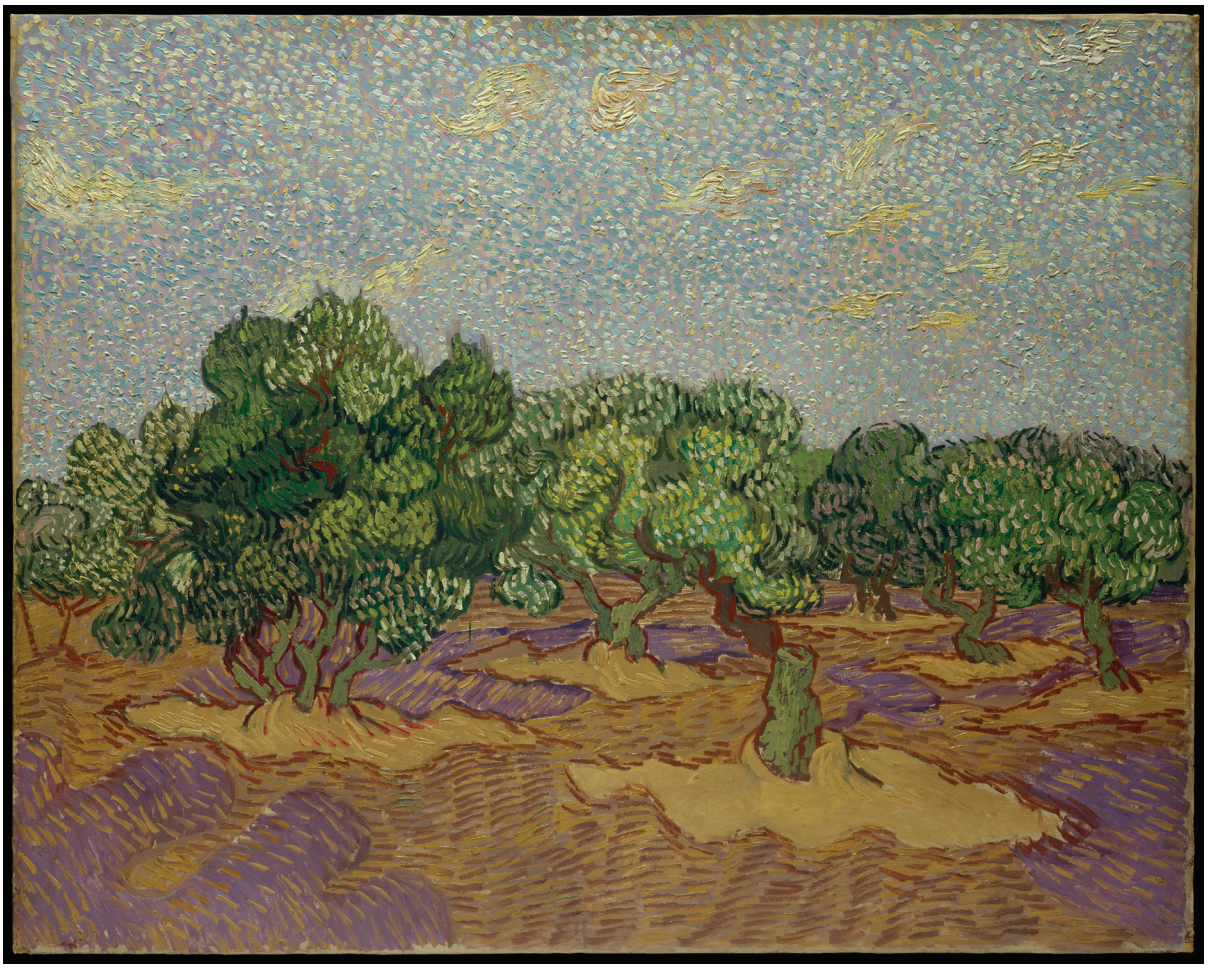

Figure 3. Vincent van Gogh, *Olive Trees*, 1889. Oil on canvas, 78 x 92 cm. The Metropolitan Museum of Art, New York, The Walter H. and Leonore Annenberg Collection, Gift of Walter H. and Leonore Annenberg, 1998, Bequest of Walter H. Annenberg, 2002, www.metmuseum.org. Creative Commons CC0 1.0 Universal Public Domain Dedication.

Seeing Patterns Where None Exist (Apophenia)

Statistical reasoning may produce outcomes not supported by human causal reasoning or theorization. Because automated systems rely on statistics when generating results, the reasons why are often difficult or impossible to explain, even by scientific experts.[94] Facial processing technologies have perpetuated this basic mistake in consequential real-world applications and in studies of portraits.[95] Yet technologists and laypeople alike often overlook the possibility that the machine's decision is absurd or incongruous. As danah boyd and Kate Crawford point out: "Too often, Big Data enables the practice of apophenia: seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions."[96]

A case in point from a GLAM dataset is The Met's Art Explorer.[97] The purpose of Art Explorer is to create "a set of pathways through the artworks that would enable user [*sic*] to traverse the catalog and find interesting associations between the artworks."[98] It is built on Microsoft's Cognitive Search technology, which performs key tasks with The Met dataset: it generates automated visual tags, performs object recognition with the images of artworks, adds to the artwork's metadata by using Microsoft's web search engine, and proposes visually similar artworks.

The last of these tasks—suggesting visually similar artworks—may be a problematic example of claiming patterns where none exist. In one case described by The Met's staff, Art Explorer matched *Olive Trees* by Vincent van Gogh (Figure 3) with *Demons Fighting over an Animal Limb* by an unknown artist believed to have worked in India (Figure 4).[99] The AI system's decision was possibly made on the basis of colors

Figure 4. Unknown artist, *Demons Fighting Over an Animal Limb*, late 1600s. Ink, opaque watercolor, and gold on paper, 29 x 19 cm. The Metropolitan Museum of Art, New York, Gift of Doris Rubin, in memory of Harry Rubin, 1989, www.metmuseum.org. Creative Commons CC0 1.0 Universal Public Domain Dedication

and shapes, according to a published explanation by the museum's general manager of collection information, Jennie Choi: "We would not normally associate Van Gogh's work with either demons or India, but looking at both images side by side, we can see similarities in the color and shape of the trees depicted in both works."[100] Then Choi praises the machine's correlations: "This is the power of AI—it can detect patterns not readily noticed by humans."[101]

The museum's response to the machine's match between *Olive Trees* and *Demons Fighting* exemplifies a key feature of Campolo and Crawford's notion of enchanted determinism: proponents of an AI system use data abundance to forego the types of explanations that have long been widely expected in our scientific era, causality and theoretical mechanisms.[102] Instead, causal explanation and theorization are thrown out in favor of the machine's capacity to find correlations.

Though any single use of Art Explorer seems harmless, it lays groundwork for an apophenia similar to that found in computational fields wherein messy, real-world sources are forgotten and replaced by orderly, mathematical proxies. The real-world sociotechnical consequences of this include risks of cultural misrepresentation.[103] For example, Ramya Srinivasan and Kanji Uchino have shown that AI scientists may conveniently define style in a correlative way that suits their algorithm's performance, and that this inappropriate problem formulation may misrepresent genuine cultural forms.[104] They cite an example of researchers who claim that their generative AI model can create "Ukiyo-e style" images, after a style found in Japanese art, by generating images with "yellowish" colors. They used a publicly available Wikiart dataset,[105] which is composed of content linked to GLAM datasets, to build their model. However, the AI-generated images fail to represent other central features of Ukiyo-e works, and the AI developers neglect to mention that their model fails to distinguish between

Ukiyo-e paintings and their copies in woodblock prints, which were historically mass-produced on paper. Whereas Ukiyo-e prints may now appear to have a yellowish support due to aging of the paper, Ukiyo-e paintings are a different medium with its own degradation and preservation issues. What this example shows, and what Parry has earlier articulated, is that GLAMs must confront the prospect of negotiating an unpredictable set of integrations for their digital content, and of contributing (even unknowingly) to AI-driven narratives that might misrepresent cultural histories.[106]

By considering how a cultural dataset can be transformed into a playground for faulty reasoning and cultural misrepresentation, GLAMs might well choose to be explicit about their limitations and set some guardrails as to how their assets may be used in the best interest of the institution and its public.

Dataset Management

Putting a dataset on the internet arguably comes with its own responsibility measures for GLAMs, which abide by institutional norms for preservation of their physical collections that need to expand to their digital assets. The unbound nature of stewarding the collection is a meaningful point of convergence with memory institutions. The importance of rigorous data management has been identified as a key intervention to mitigate harm and promote accountability within AI.[107] Critical researchers argue that datasets are long-term processes rather than static things,[108] and that dataset management is perpetual work,[109] giving rise to the imperative that organizations create dataset maintenance plans.

In an age when GLAMs undertake digital projects that connect to the public internet, dataset maintenance plans ought to account for potential uses of the dataset by others. For example, GLAM datasets may include digitized photographs of people, many of whom are no longer living, whose images carry cultural or familial meaning that may be misappropriated by certain dataset uses. AI developers have scraped the web for images of faces without the subjects' or their families' consent,[110] and the resulting datasets have been used to build surveillance systems embedded with facial processing technology.[111] In cases like these, observers have noted that well intentioned licenses for the images, such as Creative Commons licenses, fail to protect the subjects of these images from their likeness being used to develop technical systems that can be understood as repressive in certain contexts (such as authoritarian societies) and against certain groups (namely systematically marginalized populations). The Creative Commons organization's blog emphasizes several related shortfalls: "CC [Creative Commons] licenses were designed to address a specific constraint, which they do very well: unlocking restrictive copyright. But copyright is not a good tool to protect individual privacy, to address research ethics in AI development, or to regulate the use of surveillance tools employed online."[112]

Recommendation 4: Establish a dataset maintenance plan, including careful assessment of terms and conditions of use informed by the impact assessment that accounts for potential uses of the dataset by others.

GLAMs ought to consider restricting certain uses of the dataset with legal terms such as a Terms and Conditions agreement or end user license agreement.[113] GLAMs might well plan to trace dataset use by requiring that individuals register before receiving permission to download. For example, the National Palace Museum of Taiwan requires individuals to apply for an API key before using its Open API services.[114]

Dataset Security

Protecting the dataset is crucial to GLAMs' stewardship of cultural assets, and also to organization-wide security. Bad actors opposed to the mission of GLAMs might seek to attack their information technology systems to gain a ransom, disrupt operations of high-profile institutions, or gain access to systems of other organizations with which GLAMs work. Attackers might use datasets as points of entry for unauthorized access to systems, networks, or data related to donors, customers, employees, finances, operations, insurance, intellectual property, and more.[115]

Cybercriminals have realized that nonprofit organizations are terrific targets: many hold sensitive personal or intellectual property data, yet a significant portion have not conducted a basic risk assessment to identify vulnerabilities in their IT systems.[116] Cybercriminals know that an organization will pay to recover compromised data or to avoid reputational damage. Their motivations can be completely unrelated to the data's economic value.

GLAMs have to make careful decisions about datasets connected to the internet: how to store and protect the data, but also who has access to it. For example, an insecure

API can be an easy target for attackers to obtain data, gain unauthorized access to computers and networks, and/or introduce malicious data into the API.

Recommendation 5: Emphasize the institution's stewardship of the data by publicly communicating its commitment to cybersecurity.

Here are some things that GLAMs can do to embed their datasets in a public commitment to cybersecurity:

Recommendation 5a: GLAMs' datasets ought to be evaluated and secured as part of the organization's adherence to an internationally recognized security framework. A generally accepted gold standard is the U.S. National Institute of Standards and Technology (NIST) Cybersecurity Framework.[117] *The NIST framework provides risk assessment guidelines and is intended to scale with the organization's resources.*

Recommendation 5b: Get help from an organization in civil society or academia that helps nonprofits build capacity to defend against digital threats. Examples include CyberPeace Builders,[118] *the Consortium of Cybersecurity Clinics,*[119] *and Microsoft Security Program for Nonprofits.*[120]

Conclusion

The design, deployment, and monitoring of GLAM datasets are political, cultural, social, and epistemic interventions. GLAMs have an opportunity to respond to critical technology discourse by being forthright and accountable about their decisions, and by offering alternatives to commercial paradigms that emphasize dataset development at speed and scale. We have suggested that, when GLAMs critically examine their own dataset development activities, they open up more possibilities for transparency through public-facing communication and external assessment. We offered the following five recommendations toward improving GLAMs' critical reflection and accountability:

1. Conduct an impact assessment about interaction between dataset decision choices and knowledge paradigms or biases.
2. Produce comprehensive documentation, including decision provenance, of how the dataset is made. Make the documentation available for review outside the institution.
3. Consider future opportunities and risks presented by the kinds of thinking that computational methods promote, and how such futures might advance or degrade the institution's mission. Foster opportunities and mitigate risks accordingly.
4. Establish a dataset maintenance plan, including careful assessment of terms and conditions of use informed by the impact assessment, that accounts for potential uses of the dataset by others.
5. Emphasize the institution's stewardship of the data by publicly communicating its commitment to cybersecurity.

Our proposal is a call to action, focusing on the role that GLAMs can play in addressing sociotechnical problems. This set of recommendations is also an invitation for GLAMs to build consensus around how to implement better accountability mechanisms. In particular, we stress that our recommendations 1 and 2 will require alignment among GLAMs about what good impact assessment and dataset documentation look like. Toward that objective, we provided some ideas to explore:

- In the impact assessment and public-facing documentation, include information about consultation with domain experts and stakeholder groups to explore how classifications affect user experience and community representation.
- In the impact assessment and public-facing documentation, record assumptions and procedures for data cleaning.
- If annotation will rely on crowdsourced labor, document why in the impact assessment and public-facing documentation, and create a record of the annotation process, including sociodemographic information about annotators, in order to situate the resulting dataset.

There is no tidy formula for responsible dataset development in any sector. Memory institutions that take the lead on building better accountability would not only have a chance to respond to existing technology discourse and practice—they would also promote GLAMs' leadership on sociotechnical problems in the age of AI that affect public service, industry, academic, and government organizations alike. We urge memory institutions to go beyond the familiar question "What can AI do for GLAMs?" by pursuing the more underappreciated inquiry, "What can GLAMs do for AI?"

NOTES

1 We would like to thank the following people for helpful feedback on earlier drafts of this article: Amelia Saul, Hanlin Li, Jessica Newman, and Negar Rostamzadeh.

2 See, e.g., "Art + Data: Building the SFMOMA Collection API," MW2015: Museums and the Web 2015, last modified January 30, 2015, https://mw2015.museumsandtheweb.com/paper/art-data-building-the-sfmoma-collection-api/.

3 Maria Kessler, "The Met x Microsoft x MIT," Metropolitan Museum of Art, last modified February 21, 2019, https://www.metmuseum.org/blogs/now-at-the-met/2019/met-microsoft-mit-reveal-event-video.

4 "*Art Project 2023*," João Enxuto and Erica Love, performed January 14, 2014, https://theoriginalcopy.net/art-project-2023; Kimberly Christen, "Relationships, Not Records: Digital Heritage and the Ethics of Sharing Indigenous Knowledge Online," in *Routledge Companion to Media Studies and Digital Humanities*, edited by Jentery Sayers, (Routledge: Taylor and Francis, 2018), 403–12; Nanna Bonde Thylstrup, *The Politics of Mass Digitization* (Cambridge: MIT Press, 2018); Thomas Padilla, "Responsible Operations: Data Science, Machine Learning, and AI in Libraries," (Dublin, OH: OCLC Research, 2019), https://doi.org/10.25333/xk7z-9g97; Daniela Agostinho, "Care," in *Uncertain Archives: Critical Keywords for Big Data*, edited by Nanna Bonde Thylstrup, Daniela Agostinho, Annie Ring, Catherine D'Ignazio, and Kristin Veel (Cambridge: MIT Press, 2021), 75–86; Lucas Nunes Sequeira, Rafael Tsuha, et al., "A Crack Within the Museum: Problematizing Computer Vision of Commercial AIs," 2020, https://sites.usp.br/gaia/wp-content/uploads/sites/719/2020/08/zine1.pdf.

5 Ross Parry, "Transfer Protocols: Museum Codes and Ethics in the New Digital Environment," in *The Routledge Companion to Museum Ethics: Redefining Ethics for the Twenty-first Century Museum*, edited by Janet Marstine (New York: Taylor & Francis Group, 2011), 318.

6 Parry, "Transfer Protocols," 318.

7 Although The Met does not offer a public record of specific processes and personnel that contributed to its dataset creation process, to our knowledge, we may draw certain inferences. Throughout this article, we base our case study analysis in examination of the dataset itself and documents published on the web by the museum. See the announcement of the Open Access initiative in "Open Access at The Met," The Metropolitan Museum of Art, n.d., https://www.metmuseum.org/about-the-met/policies-and-documents/open-access, accessed January 29, 2022.

8 "The Metropolitan Museum of Art Collection API," Metropolitan Museum of Art and GitHub, last modified November 17, 2020, https://metmuseum.github.io/, "The Metropolitan Museum of Art Open Access CSV," Metropolitan Museum of Art and GitHub, last modified March 7, 2022, https://github.com/metmuseum/openaccess.

9 Bernard Koch, Emily Denton, Alex Hanna, Jacob G. Foster. "Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research," in *35th Conference on Neural Information Processing Systems (NeurIPS)*, (Sydney: 2021), https://doi.org/10.48550/arXiv.2112.01716.

10 Cf. the case study on The Met in Oonagh Murphy and Elena Villaespesa, "AI: A Museum Planning Toolkit," report for The Museums + AI Network (London: Goldsmiths, University of London, January 2020), 8–9, https://research.gold.ac.uk/id/eprint/28201/.

11 "Open Definition," Open Knowledge Foundation, n.d., https://opendefinition.org/, accessed May 28, 2022.

12 Douglas McCarthy and Andrea Wallace, "Open GLAM Survey Backup," Internet Archive, last modified February 17, 2022, https://archive.org/details/OpenGLAM_Survey_20220217.

13 "About Us," Europeana, n.d., https://www.europeana.eu/en/about-us, accessed April 3, 2022.

14 "Open Content Program," J. Paul Getty Museum, n.d., https://www.getty.edu/about/whatwedo/opencontent.html, accessed April 3, 2022.

15 "Open Data," National Palace Museum, n.d., https://theme.npm.edu.tw/opendata/?lang=2, accessed April 3, 2022.

16 "Open Access," Cleveland Museum of Art, n.d., https://www.clevelandart.org/open-access, accessed April 3, 2022.

17 "Open Access," Art Institute of Chicago, n.d., https://www.artic.edu/open-access, accessed April 3, 2022.

18 "Smithsonian Open Access," Smithsonian Institution, n.d., https://www.si.edu/openaccess, accessed April 3, 2022.

19 "Open Access at the National Gallery of Art," National Gallery of Art, n.d., https://www.nga.gov/open-access-images.html, accessed April 3, 2022.

20 Museum of Applied Arts and Sciences, "MAAS API Documentation," https://api.maas.museum/docs, accessed August 27, 2022.

21 "Open Access at The Met," Metropolitan Museum of Art.

22 "The Santa Barbara Statement on Collections as Data, Version 2," Always Already Computational - Collections as Data, https://collectionsasdata.github.io/statement/, accessed March 15, 2023.

23 "Partner With Us," Google Arts & Culture, n.d., https://about.artsandculture.google.com/partners/, accessed April 3, 2022.

24 "Microsoft Corporation," Cleveland Museum of Art, n.d., https://www.clevelandart.org/microsoft-corporation, accessed April 17, 2022.

25 "Smithsonian Open Access," Smithsonian Institution.

26 "Artsy," Cleveland Museum of Art, n.d., https://www.clevelandart.org/artsy, accessed April 17, 2022.

27 "The National Gallery of Art on Wikimedia Commons and Wikidata," National Gallery of Art, n.d., https://www.nga.gov/open-access-images/wikimedia-commons-wikidata.html, accessed April 17, 2022.

28 "Where the World Builds Software," GitHub, n.d., https://github.com/.

29 "(In)visible Artifacts: Parsons Students Explore the Smithsonian Collections with Online Data Visualization Projects," Smithsonian Institution, last modified February 25, 2021, https://www.si.edu/openaccess/updates/parsons-visualizations.

30 "Open Access," Cleveland Museum of Art.

31 "Europeana Aggregators," Europeana, n.d., https://pro.europeana.eu/page/aggregators, accessed April 3, 2022.

32 "Directory of Crowdsourcing Projects," Non-Profit Crowd, n.d., http://nonprofitcrowd.org/crowdsourcing-website-directory/, accessed April 3, 2022.

33 Kessler, "The Met x Microsoft x MIT."

34 Kessler, "The Met x Microsoft x MIT."

35 "Show Your Work: Parsons Students Design Stunning Data Visualizations with Met Open Access API," Metropolitan Museum of Art, last modified February 7, 2020, https://www.metmuseum.org/blogs/collection-insights/2020/met-api-parsons-data-visualization.

36 Sofie Andersen and Spencer Kiser, "Celebrating Three Years of Open Access at The Met," Metropolitan Museum of Art, last modified February 19, 2020, https://www.metmuseum.org/blogs/collection-insights/2020/met-api-third-anniversary.

37 "The Metropolitan Museum of Art Collection API," Metropolitan Museum of Art and GitHub.

38 "The Metropolitan Museum of Art Open Access CSV," Metropolitan Museum of Art and GitHub.

39 Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer, "The ArtBench Dataset: Benchmarking Generative Models with Artworks," last modified June 22, 2022, https://arxiv.org/pdf/2206.11404.pdf.

40 "Terms and Conditions/Terms of Use," Metropolitan Muse-

um of Art, last modified October 25, 2018, https://www.metmuseum.org/information/terms-and-conditions.

41 A publicly available tool suggests that images from The Met dataset made it into Stable Diffusion through its primary training dataset LAION-5B, which draws on images from The Met through Pinterest and Wikimedia. The tool does not show that anyone took The Met's dataset wholesale for training Stable Diffusion, but it does show that image and text data for individual museum objects from The Met's dataset contribute to Stable Diffusion. The tool is described by the creator in Andy Baio, "Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator," last modified August 30, 2022, https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/. The tool is discussed in James Vincent, "Anyone Can Use this AI Art Generator—That's the Risk," *The Verge, September 15, 2022, https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data.*

42 Quoted in Sindhu Sundar, "OpenAI CEO Says it's not 'A Big Dunk' that He Fears Super Intelligent AI, and its Risks are 'Far Beyond Anything We're Prepared For,'" *Business Insider*, March 27, 2023, https://www.businessinsider.com/openai-ceo-sam-altman-comments-ai-fears-risks-artificial-intelligence-2023-3.

43 Parry, "Transfer Protocols," 318.

44 Alex Hanna, Emily Denton, Andrew Smart, Hilary Nicole, and Razvan Amironesei, "Lines of Sight," *Logic 12, Commons,* December 16, 2020, accessed July 7, 2022, https://logicmag.io/commons/lines-of-sight; Milagros Miceli, Martin Schuessler, and Tianling Yang, "Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision," in *Proceedings of the ACM on Human-Computer Interaction* 4 (New York: Association for Computing Machinery, October 2020), https://dl.acm.org/doi/10.1145/3415186; Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna, "Data and its (D)iscontents: A Survey of Dataset Development and Use in Machine Learning Research," *Patterns* 2, no. 11 (2021): 4, https://doi.org/10.1016/j.patter.2021.100336; Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry, "From ImageNet to Image Classification: Contextualizing Progress on Benchmarks," in *Proceedings of the 37th International Conference on Machine Learning* 119 (n.p.: ML Research Press, 2020): 9625–35, http://proceedings.mlr.press/v119/tsipras20a.html; Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole, "On the Geneaology of Machine Learning Datasets: A Critical History of ImageNet," *Big Data & Society* 8, no. 2 (2021): 1–14, https://doi.org/10.1177/20539517211035955; Inioluwa Deborah Raji and Genevieve Fried, "About Face: A Survey of Facial Recognition Evaluation," in *Association for the Advancement of Artificial Intelligence 2020 Workshop on AI Evaluation* (Palo Alto: Association for the Advancement of Artificial Intelligence, 2021), https://arxiv.org/pdf/2102.00813.pdf; Kate Crawford, *Atlas of AI* (New Haven: Yale University Press, 2021); Eun Seo Jo and Timnit Gebru, "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning", in *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency* (FAT* '20) (New York: Association for Computing Machinery, 2020), 306–16, https://doi.org/10.1145/3351095.3372829; Kenny Peng, Arunesh Mathur, and Arvind Narayanan, "Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers," in *Proceedings of Neural Information Processing Systems* (NeurIPS 2021) Track on Datasets and Benchmarks (Sydney: 2021), https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/077e29b11be80ab57e1a2ecabb7da330-Paper-round2.pdf.

45 Ortwin Renn and Klaus Lucas, "Systemic Risk: The Threat to Societal Diversity and Coherence," *Risk Analysis 42, no. 9 (2022): 2, https://doi.org/10.1111/risa.13654.*

46 E.g., Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley, "No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World," in *31st Conference on Neural Information Processing Systems (NIPS) (Long Beach: 2017), https://research.google/pubs/pub46553/; Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten, "Does Object Recognition Work for Everyone?" in Conference on Computer Vision and Pattern Recognition (CVPR) (Long Beach: 2019), https://doi.org/10.48550/arXiv.1906.02659; Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in Conference on Fairness, Accountability, and Transparency, Journal of Machine Learning Research 81 (2018): 1–15, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.*

47 Rebecca Heilweil, "The World's Scariest Facial Recognition Company, Explained," Vox Recode (updated May 8, 2020), accessed June 30, 2022, https://www.vox.com/recode/2020/2/11/21131991/clearview-ai-facial-recognition-database-law-enforcement.

48 See "Context Stripping" below, pp. 13–14.

49 See below, nn. 51–53, 55, 57, 62–65, 74–80, 83, 86–95, 102–04, 107–12, and accompanying text.

50 Thylstrup, The Politics of Mass Digitization, 3–33.

51 Sasha Costanza-Chock, "Design Justice, A.I., and Escape from the Matrix of Domination," Journal of Design and Science (July 16, 2018), accessed July 7, 2022, https://doi.org/10.21428/96c8d426.

52 Ramya Srinivasan, Emily Denton, Jordan Famularo, Negar Rostamzadeh, Fernando Diaz, and Beth Coleman, "Artsheets for Art Datasets," in *Proceedings of Neural Information Processing Systems (NeurIPS 2021), Track on Datasets and Benchmarks (Sydney: 2021), https://research.google/pubs/pub51056/.*

53 Kinks Izsak et al., "Opportunities and Challenges of AI Technologies for the Cultural and Creative Sectors," (Luxembourg: Publications Office of the European Union, February 2022), 39, https://creative-europe.lu/publication/opportunities-and-challenges-of-ai-technologies-for-the-cultural-and-creative-sectors/.

54 Janet Marstine, "The Contingent Nature of New Museum Ethics," in *The Routledge Companion to Museum Ethics: Redefining Ethics for the Twenty-first Century Museum, edited by Janet Marstine (London and New York: Taylor & Francis Group, 2011), 14.*

55 Zohar Efroni, "The Digital Services Act: Risk-based Regulation of Online Platforms," *Internet Policy Review Opinion (Nov. 16, 2021), https://policyreview.info/articles/news/digital-services-act-risk-based-regulation-online-platforms/1606.*

56 Ulrich Beck, *Risk Society: Towards a New Modernity, trans.* Mark Ritter (London: Sage, 1992); Anthony Giddens, "Risk and Responsibility," *The Modern Law Review* 62, no. 1 (1999): 1–10.

57 "AI Risk Management Framework," National Institute of Standards and Technology, last modified January 26, 2023, https://www.nist.gov/itl/ai-risk-management-framework.

58 Efroni, "The Digital Services Act: Risk-based Regulation of Online Platforms."

59 David Wright, "A Framework for the Ethical Impact Assessment of Information Technology," *Ethics and Information Technology* 13 (2011): 199–226, https://doi.org/10.1007/s10676-010-9242-6.

60 See below nn. 61–65.

61 Shawn Averkamp, "Data Packaging Guide," May 14, 2018, https://github.com/saverkamp/beyond-open-data, accessed March 15, 2023; Sophie L. Ziegler, "Open Data in Cultural Heritage Institutions: Can We Be Better Than Data Brokers?" *DHQ: Digital Humanities Quarterly 14, no. 2 (2020), http://www.digitalhumanities.org/dhq/vol/14/2/000462/000462.html.*

62 E.g., Emily M. Bender and Batya Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational Linguistics 6* (2018): 587–604,

https://doi.org/10.1162/tacl_a_00041; Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani, "Interactive Model Cards: A Human-Centered Approach to Model Documentation," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (New York: Association for Computing Machinery, 2022), 427–39, https://doi.org/10.1145/3531146.3533108; Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson, "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT) (New York: Association for Computing Machinery, 2022), 1776–1826, https://doi.org/10.1145/3531146.3533231.*

63 Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, "Model Cards for Model Reporting," in Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*) (Atlanta: 2019), 220–229, https://doi.org/10.1145/3287560.3287596.

64 Bender and Friedman, "Data Statements for Natural Language Processing"; Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford, "Datasheets for Datasets," in *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (Stockholm: 2018),* https://doi.org/10.48550/arXiv.1803.09010; Eun Seo Jo and Timnit Gebru, "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning", in *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency (FAT* '20) (New York: Association for Computing Machinery, 2020), 306–16,* https://doi.org/10.1145/3351095.3372829; Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell, "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (New York: Association for Computing Machinery, 2021), 560–75,* https://doi.org/10.1145/3442188.3445918; Pushkarna, Zaldivar, and Kjartansson, "Data Cards"; Srinivasan, Denton, Famularo, Rostamzadeh, Diaz, and Coleman, "Artsheets for Art Datasets"; "The Data Nutrition Project," last updated 2021, https://datanutrition.org/, accessed July 23, 2022.

65 Srinivasan, Denton, Famularo, Rostamzadeh, Diaz, and Coleman, "Artsheets for Art Datasets."

66 David E. Pozen, "Seeing Transparency More Clearly," *Public Administration Review 80, no. 2 (2019): 326–31.*

67 Email correspondence with Julie Zeftel, Senior Manager of Rights and Permissions, Metropolitan Museum of Art (Oct. 4, 2021).

68 Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences* (Cambridge: MIT Press, 1999).

69 The CSV file presents 11 categories about the artist: Artist Role, Artist Prefix, Artist Display Name, Artist Display Bio, Artist Suffix, Artist Alpha Sort, Artist Nationality, Artist Begin Date, Artist End Date, Artist Gender, Artist ULAN URL, Artist Wikidata URL.

70 On Big Data's tendencies to constrain inquiry see danah boyd and Kate Crawford, "Critical Questions for Big Data," *Information, Communication & Society* 15, no. 5 (2012): 666 ["We must ask difficult questions of Big Data's models of intelligibility before they crystallize into new orthodoxies. . . . [T]he specialized tools of Big Data also have their own inbuilt limitations and restrictions. For example, Twitter and Facebook are examples of Big Data sources that offer very poor archiving and search functions. Consequently, researchers are much more likely to focus on something in the present or immediate past . . . because of the sheer difficulty or impossibility of accessing older data."]

71 Of course, the sculpture's community of origin would have used its own language(s) and terms to refer to this object in lieu of a single English title.

72 Martin Lerner, "Seated Jain Tirthankara," in *Recent Acquisitions: A Selection 1992–1993*, *The Metropolitan Museum of Art Bulletin* 51, no. 2 (1993): 92.

73 boyd and Crawford, "Critical Questions for Big Data," 667.

74 Anja Bechmann and Bender Zevenbergen, "AI and Machine Learning: Internet Research
Ethics Guidelines," Companion 6.1 in *Internet Research: Ethical Guidelines 3.0*, ed. aline shakti franzke, Anja Bechmann, Michael Zimmer, and Charles Ess, (n.p.: Association of Internet Researchers, 2020), 42, https://aoir.org/reports/ethics3.pdf.

75 Bogdan Batrinca and Philip C. Treleaven, "Social Media Analytics: A Survey of Techniques, Tools and Platforms," *AI & Society* 30 (2015): 101.

76 Ryan Cordell, "Machine Learning + Libraries: A Report on the State of the Field," (Washington: Library of Congress, July 14, 2020), 53, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig.

77 Nitesh Goyal, Ian Kivlichan, Rachel Rosen, Lucy Vasserman, "Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation," in *The 25th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)* (2022), https://arxiv.org/pdf/2205.00501.pdf; Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen, "Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation," in *NeurIPS Data-Centric AI Workshop* (December 14, 2021), https://arxiv.org/abs/2112.04554; Razvan Amironesei, Dylan Baker, Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran and Rachel Rosen, "CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation," in *ACM Conference on Fairness Accountability and Transparency (*FAccT), (Seoul: 2022), 2342–51, https://doi.org/10.1145/3531146.3534647.

78 Kate Crawford and Trevor Paglen, "Excavating AI: The Politics of Training Sets for Machine Learning," Excavating AI / AI Now Institute, last modified September 19, 2019, https://excavating.ai/.

79 Mary L. Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Boston: Houghton Mifflin Harcourt, 2019).

80 Li Yuan, "How Cheap Labor Drives China's A.I. Ambitions," *New York Times*, November 25, 2018; Mary L. Gray and Siddharth Suri, "The Humans Working behind the AI Curtain," *Harvard Business Review*, January 9, 2017; Janine Berg, "Income Security in the On-demand Economy: Findings and Policy Lessons from a Survey of Crowdworkers," *Comparative Labor Law & Policy Journal* 37, no. 3 (2016): 543; Carlos Toxtli, Siddharth Suri, and Saiph Savage, "Quantifying the Invisible Labor in Crowd Work," in *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2, (New York: Association for Computing Machinery, October 2021), Article 319, 1–26, https://dl.acm.org/doi/abs/10.1145/3476060; Crawford, *Atlas of AI*, 63-65, with references.

81 Jennie Choi, "Exploring Art with Open Access and AI: What's Next?" Metropolitan Museum of Art, last modified June 11, 2019, https://www.metmuseum.org/blogs/now-at-the-met/2019/met-microsoft-mit-exploring-art-open-access-ai-whats-next.

82 Ibid.

83 Peter J. Denning and Matti Tedre, "Computational Thinking for Professionals," *Communications of the ACM* 64, no. 12 (Dec. 2021): 30–33, https://doi.org/10.1145/3491268.

84 Mike Pepi, "Is a Museum a Database?: Institutional Conditions in Net Utopia," *e-flux journal* 60 (2014), https://www.e-flux.com/journal/60/61026/is-a-museum-a-database-institutional-conditions-in-net-utopia/.

85 Ibid.

86 Alexander Campolo and Kate Crawford, "Enchanted Determinism: Power without Responsibility in Artificial Intelligence,"

Engaging Science, Technology, and Society 6 (2020): 1–19.

87 Kate Crawford, Kate Miltner, and Mary L. Gray, "Critiquing Big Data: Politics, Ethics, Epistemology," *International Journal of Communication* 8 (2014): 1663–72; Lisa Gitelman, ed., *"Raw Data" Is an Oxymoron* (Cambridge: MIT Press, 2013); Catherine D'Ignazio and Lauren F. Klein, *Data Feminism* (Cambidge: MIT Press, 2020).

88 Louise Amoore, *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others* (Durham: Duke University Press, 2020), 16: "to consider algorithms as instruments of perception is to appreciate the processes of feature extraction, reduction, and condensation through which algorithms generate what is of interest in the data environment. . . . A defining ethical problem of the algorithm concerns not primarily the power to see, to collect, or to survey a vast data landscape, but the power to perceive and distill something for action. Algorithms function with something like an aperture—an opening that is simultaneously a narrowing, a closure, and an opening onto a scene."

89 Campolo and Crawford, ""Enchanted Determinism," 10.

90 Abigail Z. Jacobs and Hanna Wallach, "Measurement and Fairness," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (New York: Association for Computing Machinery, 2021), 375–85, https://doi.org/10.1145/3442188.3445901.

91 Annette N. Markham, "Afterword: Ethics as Impact—Moving From Error-Avoidance and Concept-Driven Models to a Future-Oriented Approach," *Social Media + Society* 4, no. 3 (2018): n.p.; boyd and Crawford, "Critical Questions for Big Data," 670; Joanna Radin, "'Digital Natives': How Medical and Indigenous Histories Matter for Big Data," *Osiris* 32, no. 1 (2017): 43–64.

92 Vinodkumar Prabhakaran, Rida Qadri and Ben Hutchinson, "Cultural Incongruencies in Artificial Intelligence," in *Proceedings of Neural Information Processing Systems (NeurIPS), Workshop on Cultures in AI/AI in Culture (2022), https://arxiv.org/abs/2211.13069; Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk, "Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction," last modified February 9, 2023,* https://arxiv.org/abs/2210.05791.

93 Christopher Nygren and Sonja Drimmer, "Art History and AI: Ten Axioms," International Journal for Digital Art History 9 (2023): 5.02–5.13, https://doi.org/10.11588/dah.2023.9.90400.

94 Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data. A Revolution that will Transform How We Live, Work, and Think* (New York: Houghton Mifflin Harcourt, 2013), 13; Crawford, *Atlas of AI*, 213, citing Peter Galison, "The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision," *Critical Inquiry* 21, no. 1 (1994): 228–66; David J. Leinweber, "Stupid Data Miner Tricks: Overfitting the S&P 500," *The Journal of Investing* 16, no. 1 (2007): 15–22; Sara Beery, Grant Van Horn, and Pietro Perona, "Recognition in Terra Incognita," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Cham: Springer, 2018), 472–89; Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel, "ImageNet-trained Cnns are Biased Toward Texture; Increasing Shape Bias Improves Accuracy and Robustness," in *International Conference on Learning Representations* (New Orleans: 2018), https://doi.org/10.48550/arXiv.1811.12231; Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, "Adversarial Examples are Not Bugs, They Are Features," in *Advances in Neural Information Processing Systems* 32 (2019): 125–36.

95 Yael Rice and Sonja Drimmer, "How Scientists Use and Abuse Portraiture," *Hyperallergic* (December 11, 2020), accessed Dec. 13, 2020, https://hyperallergic.com/604897/how-scientists-use-and-abuse-portraiture/.

96 boyd and Crawford, "Critical Questions for Big Data," 668.

97 "Art Explorer Powered By Cognitive Search," Metropolitan Museum of Art, last modified March 19, 2021, https://art-explorer.azurewebsites.net/search.

98 "How We Built the Art Explorer," Metropolitan Museum of Art, n.d., accessed March 19, 2021, https://art-explorer.azurewebsites.net/about.

99 Choi, "Exploring Art with Open Access and AI."

100 Ibid.

101 Ibid.

102 Campolo and Crawford, "Enchanted Determinism," 7, citing Mayer-Schoberger and Cukier, *Big Data. A Revolution*, 14; Peter V. Coveney, Edward R. Dougherty, and Roger R. Highfield. "Big Data Need Big Theory Too." *Philosophical Transactions of the Royal Society A* 374 (2016): 20160153; Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, June 23, 2008, accessed June 30, 2022, https://www.wired.com/2008/06/pb-theory; Alon Halevy, Peter Norvig, and Fernando Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems* 24 (2009): 8–12.

103 Katharina Burgdorf, Negar Rostamzadeh, Ramya Srinivasan, and Jennifer Lena, "Looking at Creative ML Blindspots with a Sociological Lens," in CVPR workshop, *Ethical Considerations in Creative Applications of Computer Vision* (2022), https://doi.org/10.48550/arXiv.2205.13683.

104 Ramya Srinivasan and Kanji Uchino, "Biases in Generative Art— A Causal Look from the Lens of Art History," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (New York: Association for Computing Machinery, 2021), 41–51, https://doi.org/10.1145/3442188.3445869*.

105 WikiArt Visual Encyclopedia, "About," http://www.wikiart.org/en/about, accessed March 14, 2023.

106 Parry, "Transfer Protocols," 327.

107 Kenny Peng, Arunesh Mathur, and Arvind Narayanan, "Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers," in *Proceedings of Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks* (Sydney: 2021), https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/077e29b11be80ab57e1a2ecabb7da330-Paper-round2.pdf; Hutchinson, Smart, Hanna, Denton, Greer, Kjartansson, Barnes, and Mitchell, "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure"; Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna, "Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development," in *Proceedings of the ACM on Human-Computer Interaction* 5 (New York: Association for Computing Machinery, 2021), Article 317, 1–37; Alexandra S. Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford, "A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, (Seoul: 2022), 199–212, https://doi.org/10.1145/3531146.3533086.

108 Jordan Famularo, Betty Hensellek, and Philip Walsh, "Data Stewardship: A Letter to Computer Vision from Cultural Heritage Studies," CVPR workshop, Beyond Fairness: Towards a Just, Equitable and Accountable Computer Vision, (June 25, 2021), https://sites.google.com/view/beyond-fairness-cv/accepted-papers?authuser=0.

109 Vinay Uday Prabhu and Abeba Birhane, "Large Datasets: A Pyrrhic Win for Computer Vision?" arXiv preprint, last modified July 24, 2020, https://arxiv.org/abs/2006.16923; Hutchinson, Smart, Hanna, Denton, Greer, Kjartansson, Barnes, and Mitchell, "Towards Accountability for Machine Learning Datasets"; Famularo, Hensellek, and Walsh, "Data Stewardship: A Letter to Computer Vision from Cultural Heritage Studies."

110 Prabhu and Birhane, "Large Datasets: A Pyrrhic Win for Computer Vision?"

111 Olivia Solon, "Facial Recognition's 'Dirty Little Secret': Millions of Online Photos Scraped without Consent," NBC News, March

12, 2019, accessed March 29, 2023, https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921/.

112 Ryan Merkley, "Use and Fair Use: Statement on Shared Images in Facial Recognition AI," Creative Commons blog, March 13, 2019, accessed March 29, 2023, https://creativecommons.org/2019/03/13/statement-on-shared-images-in-facial-recognition-ai/.

113 For an example of the latter, see "Responsible AI End User License Agreement," Responsible AI Licenses, n.d., accessed March 29, 2023, https://www.licenses.ai/enduser-license.

114 "Open API," National Palace Museum.

115 "Section 1: Why do Low-Risk Organizations Need Cybersecurity?" Citizen Clinic Cybersecurity Education Center, last modified February 2, 2019, https://www.citizenclinic.io/low-resource-organizations/section-1-why-do-low-risk-organizations-need-cybersecurity.

116 Ben Gose, "Nonprofits Are at Risk of Cyberattacks. Here's What You Need to Know," *The Chronicle of Philanthropy*, January 11, 2022, accessed January 23, 2022, https://www.philanthropy.com/article/safeguarding-nonprofit-data.

117 "Cybersecurity Framework," National Institute of Standards and Technology, n.d., accessed January 23, 2022, https://www.nist.gov/cyberframework; see also Afua Bruce, "Cybersecurity for Nonprofits: A Guide," NTEN, last modified February 26, 2020, https://www.nten.org/article/cybersecurity-for-nonprofits/.

118 https://cyberpeaceinstitute.org/cyberpeacebuilders/.

119 https://cybersecurityclinics.org/.

120 https://www.microsoft.com/en-us/nonprofits/data-security.

BIBLIOGRAPHY

"About Us." *Europeana*. N.d. https://www.europeana.eu/en/about-us.

Agostinho, Daniela. "Care." In *Uncertain Archives: Critical Keywords for Big Data*, edited by Nanna Bonde Thylstrup, Daniela Agostinho, Annie Ring, Catherine D'Ignazio, and Kristin Veel, 75–86. Cambridge: MIT Press, 2021.

"AI Risk Management Framework." *National Institute of Standards and Technology*. Last modified January 26, 2023. https://www.nist.gov/itl/ai-risk-management-framework.

Allyn, Bobby. "IBM Abandons Facial Recognition Products, Condemns Racially Biased Surveillance." *NPR*, June 9, 2020. Accessed June 30, 2022. https://www.npr.org/2020/06/09/873298837/ibm-abandons-facial-recognition-products-condemns-racially-biased-surveillance.

Amironesei, Razvan, Dylan Baker, Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran and Rachel Rosen. "CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation." In *ACM Conference on Fairness Accountability and Transparency (FAccT)*. Seoul: 2022. 2342–51. https://doi.org/10.1145/3531146.3534647.

Amoore, Louise. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham: Duke University Press, 2020.

Ananny, Mike. "Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness." *Science, Technology, & Human Values* 41, no. 1 (2016): 93–117.

Andersen, Sofie and Spencer Kiser. "Celebrating Three Years of Open Access at The Met." *Metropolitan Museum of Art*. Last modified February 19, 2020. https://www.metmuseum.org/blogs/collection-insights/2020/met-api-third-anniversary.

Anderson, Chris. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired* June 23, 2008. Accessed June 30, 2022. https://www.wired.com/2008/06/pb-theory.

"Art + Data: Building the SFMOMA Collection API." *MW2015: Museums and the Web 2015*. Last modified January 30, 2015. https://mw2015.museumsandtheweb.com/paper/art-data-building-the-sfmoma-collection-api/.

"Art Explorer Powered By Cognitive Search." *Metropolitan Museum of Art*. Last modified March 19, 2021. https://art-explorer.azurewebsites.net/search.

Art Project 2023. Performed by João Enxuto and Erica Love. January 14, 2014. https://theoriginalcopy.net/art-project-2023.

"Artsy." *Cleveland Museum of Art*. N.d. https://www.clevelandart.org/artsy.

Averkamp, Shawn. "Data Packaging Guide." May 14, 2018. https://github.com/saverkamp/beyond-open-data, accessed March 15, 2023.

Baio, Andy. "Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator." Last modified August 30, 2022. https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/.

Batrinca, Bogdan and Philip C. Treleaven. "Social Media Analytics: A Survey of Techniques, Tools and Platforms." *AI & Society* 30 (2015): 89–116.

Bechmann, Anja and Bender Zevenbergen. "AI and Machine

Learning: Internet Research Ethics Guidelines." *Companion 6.1 in Internet Research: Ethical Guidelines 3.0*, edited by aline shakti franzke, Anja Bechmann, Michael Zimmer, and Charles Ess, 33–49. N.p.: Association of Internet Researchers, 2020. https://aoir.org/reports/ethics3.pdf.

Beck, Ulrich. *Risk Society: Towards a New Modernity*. Trans. Mark Ritter. London: Sage, 1992.

Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in Terra Incognita." In *Proceedings of the European Conference on Computer Vision* (ECCV). Cham: Springer, 2018. 472–89.

Bender, Emily M. and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." T*ransactions of the Association for Computational Linguistics* 6 (2018): 587–604.

Berg, Janine. "Income Security in the On-demand Economy: Findings and Policy Lessons from a Survey of Crowdworkers." *Comparative Labor Law & Policy Journal* 37, no. 3 (2016): 543–76.

Bowker, Geoffrey C. and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. Cambridge: MIT Press, 1999.

boyd, danah and Kate Crawford. "Critical Questions for Big Data." *Information, Communication & Society* 15, no. 5 (2012): 662–79.

Bruce, Afua. "Cybersecurity for Nonprofits: A Guide." *NTEN*. Last modified February 26, 2020. https://www.nten.org/article/cybersecurity-for-nonprofits/.

Buolamwini, Joy and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Conference on Fairness, Accountability, and Transparency*, Journal of Machine Learning Research 81 (2018): 1–15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Burgdorf, Katharina, Negar Rostamzadeh, Ramya Srinivasan, and Jennifer Lena. "Looking at Creative ML Blindspots with a Sociological Lens." In *CVPR workshop, Ethical Considerations in Creative Applications of Computer Vision*. 2022. https://doi.org/10.48550/arXiv.2205.13683.

Campolo, Alexander and Kate Crawford. "Enchanted Determinism: Power without Responsibility in Artificial Intelligence." *Engaging Science, Technology, and Society* 6 (2020): 1–19.

Choi, Jennie. "Engaging the Data Science Community with Met Open Access API." *Metropolitan Museum of Art*. Last modified February 13, 2020. https://www.metmuseum.org/blogs/collection-insights/2020/met-api-computer-learning.

Choi, Jennie. "Exploring Art with Open Access and AI: What's Next?" *Metropolitan Museum of Art*. Last modified June 11, 2019. https://www.metmuseum.org/blogs/now-at-the-met/2019/met-microsoft-mit-exploring-art-open-access-ai-whats-next.

Christen, Kimberly. "Relationships, Not Records: Digital Heritage and the Ethics of Sharing Indigenous Knowledge Online." In *Routledge Companion to Media Studies and Digital Humanities*, edited by Jentery Sayers, 403–12. Routledge: Taylor and Francis, 2018.

Cordell, Ryan. "Machine Learning + Libraries: A Report on the State of the Field." Washington: Library of Congress. July 14, 2020. https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig.

Costanza-Chock, Sasha. "Design Justice, A.I., and Escape from the Matrix of Domination." *Journal of Design and Science* (July 16, 2018). Accessed July 7, 2022. url: https://doi.org/10.21428/96c8d426.

Coveney, Peter V., Edward R. Dougherty, and Roger R. Highfield. "Big Data Need Big Theory Too." *Philosophical Transactions of the Royal Society A* 374 (2016): 20160153.

Crawford, Kate. *Atlas of AI*. New Haven: Yale University Press, 2021.

Crawford, Kate, Kate Miltner, and Mary L. Gray. "Critiquing Big Data: Politics, Ethics, Epistemology." *International Journal of Communication* 8 (2014): 1663–72.

Crawford, Kate and Trevor Paglen. "Excavating AI: The Politics of Training Sets for Machine Learning." *Excavating AI / AI Now Institute*. Last modified September 19, 2019. https://excavating.ai/.

Crisan, Anamaria, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. "Interactive Model Cards: A Human-Centered Approach to Model Documentation." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT). New York: Association for Computing Machinery, 2022. 427–39. https://doi.org/10.1145/3531146.3533108.

"Cybersecurity Framework." *National Institute of Standards and Technology*. N.d. https://www.nist.gov/cyberframework.

Denning, Peter J. and Matti Tedre. "Computational Thinking for Professionals." *Communications of the ACM* 64, no. 12 (Dec. 2021): 30–33. https://doi.org/10.1145/3491268.

Denton, Emily, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. "On the Geneaology of Machine Learning Datasets: A Critical History of ImageNet." *Big Data & Society* 8, no. 2 (2021): 1–14. https://doi.org/10.1177/20539517211035955.

Denton, Emily, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. "Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation." *NeurIPS Data-Centric AI Workshop*. Last modified December 14, 2021. https://arxiv.org/abs/2112.04554.

DeVries, Terrance, Ishan Misra, Changhan Wang, and Laurens van der Maaten. "Does Object Recognition Work for Everyone?" In *Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach: 2019. https://doi.org/10.48550/arXiv.1906.02659.

D'Ignazio, Catherine and Lauren F. Klein. *Data Feminism*. Cambidge: MIT Press, 2020.

"Directory of Crowdsourcing Projects." Non-Profit Crowd. N.d. http://nonprofitcrowd.org/crowdsourcing-website-directory/.

Efroni, Zohar. "The Digital Services Act: Risk-based Regulation of Online Platforms." *Internet Policy Review Opinion*. November 16, 2021. https://policyreview.info/articles/news/digital-services-act-risk-based-regulation-online-platforms/1606.

"Europeana Aggregators." *Europeana*. N.d. https://pro.europeana.eu/page/aggregators.

Famularo, Jordan, Betty Hensellek, and Philip Walsh. "Data Stewardship: A Letter to Computer Vision from Cultural Heritage Studies." In *CVPR workshop, Beyond Fairness: Towards a Just, Equitable and Accountable Computer Vision*. June 25, 2021. https://sites.google.com/view/beyond-fairness-cv/accepted-papers?authuser=0.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for Datasets." In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*. Stockholm: 2018. https://doi.org/10.48550/arXiv.1803.09010.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. "ImageNet-trained Cnns are Biased Toward Texture; Increasing Shape Bias Improves Accuracy and Robustness." In Int*ernational Conference on Learning Representations*. New Orleans, 2018. https://doi.org/10.48550/arXiv.1811.12231.

Giddens, Anthony. "Risk and Responsibility." *The Modern Law Review* 62, no. 1 (1999): 1–10.

Gitelman, Lisa, ed. *"Raw Data" Is an Oxymoron*. Cambridge: MIT Press, 2013.

Gose, Ben. "Nonprofits Are at Risk of Cyberattacks. Here's What You Need to Know." *The Chronicle of Philanthropy*, January 11, 2022. Accessed January 23, 2022. https://www.philanthropy.com/article/safeguarding-nonprofit-data.

Goyal, Nitesh, Ian Kivlichan, Rachel Rosen, Lucy Vasserman. "Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation." In The 25th ACM Conferen*ce On Computer-Supported Cooperative Work And Social Computing (CSCW)*. Virtual, 2022. https://arxiv.org/pdf/2205.00501.pdf.

Gray, Mary L. and Siddharth Suri. Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. Boston: Houghton Mifflin Harcourt, 2019.

Gray, Mary L. and Siddharth Suri. "The Humans Working behind the AI Curtain." *Harvard Business Review*, January 9, 2017.

Halevy, Alon, Peter Norvig, and Fernando Pereira. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24 (2009): 8–12.

Hanna, Alex, Emily Denton, Andrew Smart, Hilary Nicole, and Razvan Amironesei. "Lines of Sight." *Logic 12*, Commons (December 16, 2020). Accessed July 7, 2022. https://logicmag.io/commons/lines-of-sight.

Heilweil, Rebecca. "The World's Scariest Facial Recognition Company, Explained." *Vox Recode*, updated May 8, 2020. Accessed June 30, 2022. https://www.vox.com/recode/2020/2/11/21131991/clearview-ai-facial-recognition-database-law-enforcement.

"How We Built the Art Explorer." *Metropolitan Museum of Art*. N.d. https://art-explorer.azurewebsites.net/about.

Hutchinson, Ben, Andrew Smart, Alex Hanna, Emily

Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. New York: Association for Computing Machinery, 2021. 560–75. https://doi.org/10.1145/3442188.3445918.

Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. "Adversarial Examples are Not Bugs, They Are Features." In *Advances in Neural Information Processing Systems* 32 (2019): 125–36.

"(In)visible Artifacts: Parsons Students Explore the Smithsonian Collections with Online Data Visualization Projects." *Smithsonian*. Last modified February 25, 2021. https://www.si.edu/openaccess/updates/parsons-visualizations.

Jacobs, Abigail Z. and Hanna Wallach. "Measurement and Fairness." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. New York: Association for Computing Machinery, 2021. 375–85. https://doi.org/10.1145/3442188.3445901.

Jo, Eun Seo and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." In *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency* (FAT* '20). New York: Association for Computing Machinery, 2020. 306–16. https://doi.org/10.1145/3351095.3372829.

Kessler, Maria. "The Met x Microsoft x MIT." *Metropolitan Museum of Art*. Last modified February 21, 2019. https://www.metmuseum.org/blogs/now-at-the-met/2019/met-microsoft-mit-reveal-event-video.

Koch, Bernard, Emily Denton, Alex Hanna, and Jacob G. Foster. "Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research." In *35th Conference on Neural Information Processing Systems (NeurIPS)*. Sydney: 2021. https://doi.org/10.48550/arXiv.2112.01716.

Leinweber, David J. "Stupid Data Miner Tricks: Overfitting the S&P 500." *The Journal of Investing 16*, no. 1 (2007): 15–22.

Lerner, Martin. "Seated Jain Tirthankara." In Recent Acquisitions: A Selection 1992–1993, *The Metropolitan Museum of Art Bulletin* 51, no. 2 (1993): 92.

Liao, Peiyuan, Xiuyu Li, Xihui Liu, and Kurt Keutzer. "The ArtBench Dataset: Benchmarking Generative Models with Artworks." Last modified June 22, 2022. https://arxiv.org/pdf/2206.11404.pdf.

Luccioni, Alexandra S. ,Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. "A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication." In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. (Seoul: 2022. 199–212. https://doi.org/10.1145/3531146.3533086.

"MAAS API Documentation." *Museum of Applied Arts and Sciences*. N.d. https://api.maas.museum/docs.

Markham, Annette N. "Afterword: Ethics as Impact—Moving From Error-Avoidance and Concept-Driven Models to a Future-Oriented Approach." *Social Media + Society* 4, no. 3 (2018): n.p., https://doi.org/10.1177/2056305118784504.

Marstine, Janet. "The Contingent Nature of New Museum Ethics." In *The Routledge Companion to Museum Ethics: Redefining Ethics for the Twenty-first Century Museum*. Edited by Janet Marstine. London and New York: Taylor & Francis Group, 2011. 3–25.

Mayer-Schönberger, Viktor and Kenneth Cukier. *Big Data. A Revolution that will Transform How We Live, Work, and Think*. New York: Houghton Mifflin Harcourt, 2013.

McCarthy, Douglas and Andrea Wallace. "Open GLAM Survey Backup." *Internet Archive*. Last modified February 17, 2022. https://archive.org/details/OpenGLAM_Survey_20220217.

Merkley, Ryan. "Use and Fair Use: Statement on Shared Images in Facial Recognition AI." *Creative Commons blog*. March 13, 2019. Accessed March 29, 2023. https://creativecommons.org/2019/03/13/statement-on-shared-images-in-facial-recognition-ai/.

Miceli, Milagros, Martin Schuessler, and Tianling Yang. "Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision." In *Proceedings of the ACM on Human-Computer Interaction* 4. New York: Association for Computing Machinery, October 2020. https://dl.acm.org/doi/10.1145/3415186.

"Microsoft Corporation." *Cleveland Museum of Art*. N.d. https://www.clevelandart.org/microsoft-corporation.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer,

Inioluwa Deborah Raji, and Timnit Gebru. "Model Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. Atlanta: 2019. 220–229. https://doi.org/10.1145/3287560.3287596.

Murphy, Oonagh and Elena Villaespesa. "AI: A Museum Planning Toolkit." *Report for The Museums + AI Network*. London: Goldsmiths, University of London. January 2020. https://research.gold.ac.uk/id/eprint/28201/.

Nygren, Christopher and Sonja Drimmer. "Art History and AI: Ten Axioms." *International Journal for Digital Art History* 9 (2023): 5.02–5.13. https://doi.org/10.11588/dah.2023.9.90400.

Oldman, Dominic, Diana Tanase, and Stephanie Santschi. "The Problem of Distance — A ResearchSpace Case Study on Sequencing Hokusai Print Impressions to Form a Human Curated Network of Knowledge." *International Journal for Digital Art History* 4 (2019): 5.29–5.45. https://doi.org/10.11588/dah.2019.4.72071.

O'Neil, Cathy. "How Algorithms Rule Our Working Lives." *The Guardian*, September 1, 2016. Accessed June 30, 2022. https://www.theguardian.com/science/2016/sep/01/how-algorithms-rule-our-working-lives.

"Open Access." *Art Institute of Chicago*. N.d. https://www.artic.edu/open-access.

"Open Access." *Cleveland Museum of Art*. N.d. https://www.clevelandart.org/open-access.

"Open Access at The Met." *Metropolitan Museum of Art*. N.d. https://www.metmuseum.org/about-the-met/policies-and-documents/open-access.

"Open Access at the National Gallery of Art." *National Gallery of Art*. N.d. https://www.nga.gov/open-access-images.html.

"OpenAPI." *National Palace Museum*. N.d. http://210.69.170.71/opendata/APP_Prog/eng/overview_eng.aspx.

"Open Content Program." *J. Paul Getty Museum*. N.d. https://www.getty.edu/about/whatwedo/opencontent.html.

"Open Data." *National Palace Museum*. N.d. https://theme.npm.edu.tw/opendata/?lang=2.

"Open Definition." *Open Knowledge Foundation*. N.d. https://opendefinition.org/.

Padilla, Thomas. "Responsible Operations: Data Science, Machine Learning, and AI in Libraries." Dublin, OH: OCLC Research, 2019. https://doi.org/10.25333/xk7z-9g97.

Parry, Ross. "Transfer Protocols: Museum Codes and Ethics in the New Digital Environment." In *The Routledge Companion to Museum Ethics: Redefining Ethics for the Twenty-first Century Museum*. Edited by Janet Marstine. New York: Taylor & Francis Group, 2011. 316–31.

"Partner With Us." *Google Arts & Culture*. N.d. https://about.artsandculture.google.com/partners/.

Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. "Data and its (D)iscontents: A Survey of Dataset Development and Use in Machine Learning Research." *Patterns* 2, no. 11 (2021): 1–14. https://doi.org/10.1016/j.patter.2021.100336.

Peng, Kenny, Arunesh Mathur, and Arvind Narayanan. "Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers." In *Proceedings of Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*. Sydney: 2021. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/077e29b11be80ab57e1a2ecabb7da330-Paper-round2.pdf.

Pepi, Mike. "Is a Museum a Database?: Institutional Conditions in Net Utopia." *e-flux journal* 60 (2014). https://www.e-flux.com/journal/60/61026/is-a-museum-a-database-institutional-conditions-in-net-utopia/.

Pozen, David E. "Seeing Transparency More Clearly." *Public Administration Review* 80, no. 2 (2019): 326–31.

Vinodkumar Prabhakaran, Rida Qadri and Ben Hutchinson. "Cultural Incongruencies in Artificial Intelligence." In *Proceedings of Neural Information Processing Systems (NeurIPS), Workshop on Cultures in AI/AI in Culture*. 2022. https://arxiv.org/abs/2211.13069.

Prabhu, Vinay Uday and Abeba Birhane. "Large Datasets: A Pyrrhic Win for Computer Vision?" *arXiv* preprint. Last modified July 24, 2020. https://arxiv.org/abs/2006.16923.

Pushkarna, Mahimam Andrew Zaldivar, and Oddur Kjartansson. "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI." In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. New York: Association for Computing Machinery, 2022. 1776–1826. https://doi.org/10.1145/3531146.3533231.

Radin, Joanna. "'Digital Natives': How Medical and Indigenous Histories Matter for Big Data." *Osiris* 32, no. 1 (2017): 43–64.

Raji, Inioluwa Deborah and Genevieve Fried. "About Face: A Survey of Facial Recognition Evaluation." In *Association for the Advancement of Artificial Intelligence 2020 Workshop on AI Evaluation*. Palo Alto: Association for the Advancement of Artificial Intelligence, 2021. https://arxiv.org/pdf/2102.00813.pdf.

Renn, Ortwin and Klaus Lucas. "Systemic Risk: The Threat to Societal Diversity and Coherence." *Risk Analysis 42*, no. 9 (2022): 1–14. https://doi.org/10.1111/risa.13654.

"Responsible AI End User License Agreement." *Responsible AI Licenses*. Accessed March 29, 2023. https://www.licenses.ai/enduser-license.

Rice, Yael and Sonja Drimmer. "How Scientists Use and Abuse Portraiture." *Hyperallergic*, December 11, 2020. Accessed Dec. 13, 2020. https://hyperallergic.com/604897/how-scientists-use-and-abuse-portraiture/.

"The Santa Barbara Statement on Collections as Data, Version 2." *Always Already Computational - Collections as Data*. Accessed March 15, 2023. https://collectionsasdata.github.io/statement/.

Scheuerman, Morgan Klaus, Emily Denton, and Alex Hanna. "Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development." In *Proceedings of the ACM on Human-Computer Interaction* 5. New York: Association for Computing Machinery, 2021. Article 317, 1–37.

"Section 1: Why do Low-Risk Organizations Need Cybersecurity?" *Citizen Clinic Cybersecurity Education Center*. Last modified February 2, 2019. https://www.citizenclinic.io/low-resource-organizations/section-1-why-do-low-risk-organizations-need-cybersecurity.

Sequeira, Lucas Nunes, Rafael Tsuha, et al. "A Crack Within the Museum: Problematizing Computer Vision of Commercial AIs." 2020. Accessed August 1, 2021. https://sites.usp.br/gaia/wp-content/uploads/sites/719/2020/08/zine1.pdf.

Shankar, Shreya, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. "No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World." In *31st Conference on Neural Information Processing Systems (NIPS)*. Long Beach: 2017. https://research.google/pubs/pub46553/.

Shelby, Renee, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. "Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction." Last modified February 9, 2023. https://arxiv.org/abs/2210.05791.

"Show Your Work: Parsons Students Design Stunning Data Visualizations with Met Open Access API." *Metropolitan Museum of Art*. Last modified February 7, 2020. https://www.metmuseum.org/blogs/collection-insights/2020/met-api-parsons-data-visualization.

"Smithsonian Open Access." *Smithsonian Institution*. N.d. https://www.si.edu/openaccess.

Solon, Olivia. "Facial Recognition's 'Dirty Little Secret': Millions of Online Photos Scraped without Consent." *NBC News*. March 12, 2019. Accessed March 29, 2023. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921.

Srinivasan, R., E. Denton, J. Famularo, N. Rostamzadeh, F. Diaz, and B. Coleman. "Artsheets for Art Datasets." In *Proceedings of Neural Information Processing Systems (NeurIPS 2021)*, Track on Datasets and Benchmarks. Sydney: 2021. https://research.google/pubs/pub51056/.

Srinivasan, Ramya and Kanji Uchino. "Biases in Generative Art— A Causal Look from the Lens of Art History." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT. New York: Association for Computing Machinery, 2021. 41–51. https://doi.org/10.1145/3442188.3445869.

Sundar, Sindhu. "OpenAI CEO Says it's not 'A Big Dunk' that He Fears Super Intelligent AI, and its Risks are 'Far Beyond Anything We're Prepared For.'" *Business Insider*. March 27, 2023. Accessed March 27, 2023. https://www.businessinsider.com/openai-ceo-sam-altman-comments-ai-fears-risks-artificial-intelligence-2023-3.

"Terms and Conditions/Terms of Use." *Metropolitan Museum of Art*. Last modified October 25, 2018. https://www.metmuseum.org/information/terms-and-conditions.

"The Metropolitan Museum of Art Open Access CSV." *Metropolitan Museum of Art and GitHub*. Last modified March 7, 2022. https://github.com/metmuseum/openaccess.

"The Metropolitan Museum of Art Collection API." *Metropolitan*

Museum of Art and GitHub. Last modified November 17, 2020. https://metmuseum.github.io/.

"The National Gallery of Art on Wikimedia Commons and Wikidata." *National Gallery of Art*. N.d. https://www.nga.gov/open-access-images/wikimedia-commons-wikidata.html.

Toxtli, Carlos, Siddharth Suri, and Saiph Savage. "Quantifying the Invisible Labor in Crowd Work." In *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2. New York: Association for Computing Machinery, October 2021. Article 319, 1–26. https://dl.acm.org/doi/abs/10.1145/3476060.

Thylstrup, Nanna Bonde. *The Politics of Mass Digitization*. Cambridge: MIT Press, 2018.

Tsipras, Dimitris, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. "From ImageNet to Image Classification: Contextualizing Progress on Benchmarks." In *Proceedings of the 37th International Conference on Machine Learning* 119. N.p.: ML Research Press, 2020. 9625–35. http://proceedings.mlr.press/v119/tsipras20a.html.

Vincent, James. "Anyone Can Use this AI Art Generator—That's the Risk." The Verge. September 15, 2022. https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data.

"Where the World Builds Software." *GitHub*. N.d. https://github.com/.

Wright, David. "A Framework for the Ethical Impact Assessment of Information Technology." *Ethics and Information Technology* 13 (2011): 199–226. https://doi.org/10.1007/s10676-010-9242-6.

Yuan, Li. "How Cheap Labor Drives China's A.I. Ambitions." *New York Times*, November 25, 2018.

Ziegler, Sophie L. "Open Data in Cultural Heritage Institutions: Can We Be Better Than Data Brokers?" *DHQ: Digital Humanities Quarterly* 14, no. 2 (2020). http://www.digitalhumanities.org/dhq/vol/14/2/000462/000462.html.

JORDAN FAMULARO, PhD is a consultant and former postdoctoral scholar at the University of California, Berkeley's Center for Long-Term Cybersecurity. Her cultural research develops insight into the nexus of digital harm, data curation, and social responsibility. She produces collaborative, multidisciplinary research on dataset development ethics for machine learning, which has been presented at the Conference on Computer Vision and Pattern Recognition and the Conference on Neural Information Processing Systems. Jordan received her doctorate in art history from the Institute of Fine Arts at New York University, where she specialized in technology and culture of the early modern Mediterranean world. Her work has been published in the International Journal of Corporate Social Responsibility, Brookings TechStream, Platform Governance Terminologies, Sixteenth Century Journal, and Shift. https://www.linkedin.com/in/jordan-famularo/

Correspondence email: jjf376@nyu.edu

REMI DENTON is a Staff Research Scientist at Google, within the Technology, AI, Society, and Culture team, where they study the sociocultural impacts of AI technologies and conditions of AI development. Prior to joining Google, Remi received their PhD in Computer Science from the Courant Institute of Mathematical Sciences at New York University, where they focused on unsupervised learning and generative modeling of images and video. Prior to that, they received their BSc in Computer Science and Cognitive Science at the University of Toronto. Though trained formally as a computer scientist, Remi draws ideas and methods from multiple disciplines and is drawn towards highly interdisciplinary collaborations, in order to examine AI systems from a sociotechnical perspective. Remi's recent research centers on sociocultural impacts of emerging text- and image-based generative AI, with a focus on data considerations and representational harms. https://cephaloponderer.com/wp-content/uploads/2022/10/cv.pdf

Correspondence email address: dentone@google.com