Surveillance & Society

Synthetic Data and Reverse Image Search: Constructing New Surveillant Indexicalities

Renée Ridgway and Nicolas Malevé

Volume 22, Number 4, 2024

Open Issue

URI: https://id.erudit.org/iderudit/1115678ar
DOI: https://doi.org/10.24908/ss.v22i4.18332

See table of contents

Publisher(s)

Surveillance Studies Network

ISSN

1477-7487 (digital)

Explore this journal

Cite this document

Ridgway, R. & Malevé, N. (2024). Synthetic Data and Reverse Image Search: Constructing New Surveillant Indexicalities. *Surveillance & Society*, *22*(4), 466–471. https://doi.org/10.24908/ss.v22i4.18332

Article abstract

The recent evolution of algorithmic techniques (mining, filtering, modelling) makes people more transparent through sophisticated search interactions and online monitoring, heightening opportunities for surveillance. The advent of computer-generated "synthetic data" has created another twist in the techno-information revolution of generative "artificial intelligence." Promoted by tech companies to circumvent privacy legislation and to develop cheaper monitoring technologies, synthetic data is touted as a solution to surveillance capitalism. This dialogue paper focuses on the use of synthetic data in the context of "fake" images and discriminatory technologies by first discussing the relation between representation and indexicality via the medium of (digital) photography and then via reverse image search. A digital ethnography by the authors uses artificially generated images of people "who do not exist" to query the reverse search engine PimEyes, which offers a biometric search for anyone wishing to find their faces on the internet. PimEyes finds faces similar to a person that doesn't exist, provoking questions both about the generated image used as a query and the status of the search result. The results show the tensions inherent to the use of synthetic data: a dialectic between increasing precision and increasing scepticism. When visiting the offered, linked websites, the confusion increases as the user struggles to determine if the images PimEyes found are synthetic or real. In this context, reverse image search will likely stimulate future synthetic data development and simultaneously offer services that embed metadata into files, as well as forensics to secure indexicality, introducing yet other factors into the loop between representation and generation. Therefore, the matter of concern won't be the ability to produce realistic representations through the use of synthetic data, but the demand of indexicality that their use triggers and the bureaucratic apparatuses of verification that emerge to contain it.

Surveillance & Society

Dialogue

Synthetic Data and Reverse Image Search: Constructing New Surveillant Indexicalities

Renée Ridgway

Aahrus University, Denmark
rridgway@cc.au.dk

Nicolas Malevé

Aahrus University, Denmark
maleven@cc.au.dk

Abstract

The recent evolution of algorithmic techniques (mining, filtering, modelling) makes people more transparent through sophisticated search interactions and online monitoring, heightening opportunities for surveillance. The advent of computer-generated "synthetic data" has created another twist in the techno-information revolution of generative "artificial intelligence." Promoted by tech companies to circumvent privacy legislation and to develop cheaper monitoring technologies, synthetic data is touted as a solution to surveillance capitalism. This dialogue paper focuses on the use of synthetic data in the context of "fake" images and discriminatory technologies by first discussing the relation between representation and indexicality via the medium of (digital) photography and then via reverse image search. A digital ethnography by the authors uses artificially generated images of people "who do not exist" to query the reverse search engine PimEyes, which offers a biometric search for anyone wishing to find their faces on the internet. PimEyes finds faces similar to a person that doesn't exist, provoking questions both about the generated image used as a query and the status of the search result. The results show the tensions inherent to the use of synthetic data: a dialectic between increasing precision and increasing scepticism. When visiting the offered, linked websites, the confusion increases as the user struggles to determine if the images PimEyes found are synthetic or real. In this context, reverse image search will likely stimulate future synthetic data development and simultaneously offer services that embed metadata into files, as well as forensics to secure indexicality, introducing yet other factors into the loop between representation and generation. Therefore, the matter of concern won't be the ability to produce realistic representations through the use of synthetic data, but the demand of indexicality that their use triggers and the bureaucratic apparatuses of verification that emerge to contain it.

Introduction

Generative "artificial intelligence" (AI) is now confronting citizens with unexpected novel technologies that alter their ability to engage anonymously offline and online. This recent evolution of algorithmic techniques (mining, filtering, modelling) makes people more transparent through sophisticated search interactions and monitoring by online platforms. Furthermore, the dissemination of "fake images" confuses the identification of human faces. These trends indicate the transition from a situation where one could control the exposure of their digital engagement through privacy legislation, encryption software, and/or obfuscation tactics to one of increasingly algorithmically determined publics. Individuals have become *dividuals* and "masses, samples, data, markets, or *banks*" (Deleuze 1992: 5; italics in the original) as "data doubles" (Poster 1997; Raley 2013: 127, cited by Ridgway 2021), parts of "surveillant assemblages" (Haggerty and Ericson 2000), which are constructed by "indexicality" that arrives "from elsewhere…and its regimes of objectivity" (Rouvroy 2013).

The advent of computer-generated "synthetic data," which mimics and substitutes "empirical observations without directly corresponding to real-world phenomena" (Offenhuber 2024: 1), has created another twist

Ridgway, Reneé, and Nicolas Malevé. 2024. Synthetic Data and Reverse Image Search: Constructing New Surveillant Indexicalities. *Surveillance & Society* 22 (4): 466-471.
https://ojs.library.queensu.ca/index.php/surveillance-and-society/index | ISSN: 1477-7487

in the techno-information revolution. Urban planners, Wall Street brokers, healthcare technicians, and US census organisations are modelling and predicting futures without the physical presence of objects. Promoted by big tech companies to circumvent privacy legislation and to develop cheaper tracking and monitoring technologies, synthetic data is touted as a solution to surveillance capitalism. Deemed the "new playground" for accelerating automation (Steinhoff 2022), synthetic data complicates the concept of "raw data" already put forth by critical data studies scholars who demonstrated how data are always "cooked and processed" (Bowker 2008; Gitelman 2013). Moreover, Helm, Benjamin, and Pujadas (2024: 1, 4) argue that presenting synthetic data as a fix to raw data's problems turns it into a "discursive-political device" that eludes ethical scrutiny; simultaneously, it introduces a shift in the "data economy of data collection to data production, from problems of representation to problems of design."

Increasingly, non-representational frameworks allow data to exist in multiples, as they are unexplainably generated by Artificial Neural Networks (ANNs) for different goals and do not correspond to real-world objects (Loukissas 2019; Offenhuber 2024: 6). Although synthetic data is sometimes used to protect the privacy of individuals or to decrease bias in the datasets, the ethics of what Offenhuber (2024: 13) describes as an "anything goes" attitude in Silicon Valley's world of AI now facilitates the loss of "artifacts, glitches, and data imperfections" in real images that are often entirely absent in fakes. In addition to this, the absence of the material traces of synthetic data's fabrication and its lack of data origin or provenance creates additional problems that also require scrutiny. This dialogue paper focuses on the use of synthetic data in the context of "fake" images and discriminatory technologies through reverse image search and, in particular, images that have a strong claim to indexicality.

Representation versus Indexicality

There is a long historical relationship between image techniques and indexicality, or in other words, images presenting themselves as an emanation of their referent (Barthes 1981: 80), which culminated with the invention of analogue photography.[1] As photography became a digital medium, many commentators announced a radical break as representation turned into simulation, with pixels replacing the imprint of light on film. However, this narrative has been critiqued. Daniel Rubinstein and Katrina Sluis (2008) point to the technical process of analogue photography and stress that a whole series of decisions happened in the lab— that the revealed photograph is the result of manipulations and decisions, never a simple imprint. In *The Disciplinary Frame*, John Tagg (2009) demonstrates that the photograph never stood alone in court and always needed external elements to stabilise its meaning and framing. As Allan Sekula (1986: 15) put it, the camera was never the key device of instrumental realism— the filing cabinet was.

When photographic techniques moved from analogue to digital, what occurred was not a break from an indexicality emanating from a referent to an artificial simulation but a redistribution of stabilisation techniques within the codes of representation. The consequence is a transition away from a theory of the image focusing on the photograph itself and its ontological relation to the real to one concentrating on the larger assemblages that ground the photographic. Instead of filing cabinets, today's photographic assemblages comprise environments of annotations where thousands of people tag images scraped from the web. These include the datasets on which AI models are trained, the databases and platforms responsible for the management and circulation of images, and the search engines that function as the links between these elements (Malevé 2020). If, as Rouvroy (2013) suggests**,** indexicality seems to arrive from elsewhere, it is because the relation between the image, the camera, and the engines of classification and identification have scaled up and become increasingly sophisticated.

[1] Here, indexicality is used to denote analogue photography's supposed privileged link to reality due to its ability to chemically capture light.

Furthermore, this intense process of datafication boosts algorithms' flexibility at capturing and matching patterns, disassembling the photograph into semantic units and then reassembling it: opening up the image to search—recognition as well as generation. It installs a tension in the capabilities of digital machines with "differential implications," as Louis Ravn (in this issue) notes. As these assemblages ramify exponentially, they "index" dizzying amounts of data and enable algorithms to correlate, for instance, facial patterns across billions of images, thereby increasing the potential of identification at scale. But as the same techniques significantly augment the creation of synthetic images and fakes, they instil a sense of scepticism towards any truth claim made on behalf of the image. These dialectics take place in the particulars of reverse image search.

Reverse Image Search

PimEyes is a reverse search engine offering biometric search for not just public authorities but also anyone wishing to find their own faces. After uploading an image, the user sees websites and their provenance in order to possibly delete unwanted images. Presently, its database contains more than nine-hundred million images, ostensibly scraped from publicly accessible sites on the web (Laufer and Meineck 2020; Wakefield 2020). After verifying the content, Netzpolitik's investigative journalists Laufer and Meineck questioned PimEyes about the scraping of websites, stating that GDPR legislation includes the clause "processing of biometric data for the purpose of uniquely identifying a natural person shall be prohibited." PimEyes argued that "because PimEyes does not assign names to faces, there is no legal issue," yet now it no longer allows people to upload public figures, only images of themselves. Instead, the following appears on their webpage: "Upload photo and find out where images are published" with a lighter subtext stating "Don't worry, we will not store it!" (see Figure 1).

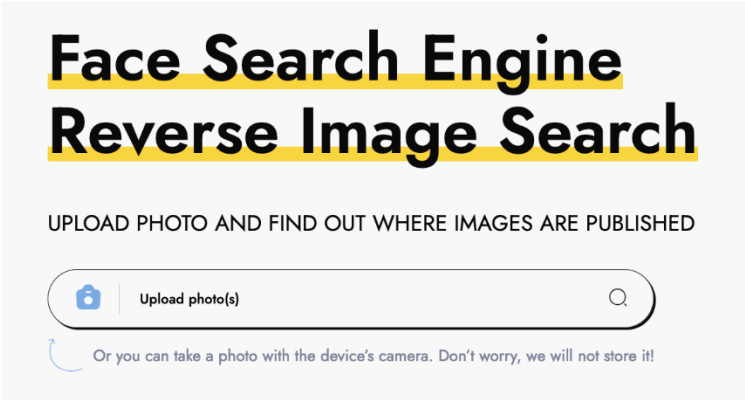

Figure 1: *Pimeyes.com screenshot of the upload form of facial reverse image search.*

A Technology in the Making

PimEyes draws on a subset of data focused on the actual identity of a person in an image, where only facial features are used for the query, in contrast to other "reverse search" engines that look for image similarity.[2] PimEyes' algorithm applies the face as a pattern independent from its surroundings, which allows the software to find a face with similar traits in images taken under very different conditions. However, it doesn't discriminate between "authentic" portraits and "fake" images if the latter are realistic enough. For instance, when carrying out a search using synthetic images generated with the website thispersondoesnotexist (2024), the results are perplexing, as they display images whose status is ambiguous

[2] TinEye (n.d.) is an example of this type of reverse image search engine.

(see Figure 2). PimEyes finds faces similar to a person that doesn't exist, provoking questions both about the generated image used as a query and the status of the results. When visiting the offered, linked websites, the confusion increases as the user struggles to determine if the images pictured are synthetic or real. At this stage of development, PimEyes is a "technology in the making," and what kind of scenarios might be available in the future is only speculation.

Figure 2: *Portraits generated on the website thispersondoesnotexist.com (2024).*

Yet, what is not speculation is that all datasets, scraped from the web or synthetically produced, contribute to the proliferation of human surveillance (Zuboff 2019), where they "help to create a market for the data that are produced and establish an exchange value for the output of surveillance" (Duke 2023). Although PimEyes has a surveillance payment model for mass searches and markets itself as a tool for "digital self-defense, so that for example users can detect possible fake profiles of themselves," (PimEyes n.d.), privacy activists and lawyers state that it is an "infringement of the right to one's own image" (Laufer and Meineck 2020). Moreover, this technology could still be abused by stalkers, intelligence services, and law enforcement to identify anyone who desires anonymity in the public sphere, as well as those who participate in the act of protesting.

Data Are Not Necessarily People After All

Offenhuber (2024: 2) shows how a departure from a representational concept of data could be obtained with a "relational model," where the phenomenon of synthetic data is defined by "their purpose, performances, and contexts of use." Applying this relational model of analysis helps highlight the purpose and performance of synthetic data in our interaction with PimEyes. Tying together photography, indexicality, objectivity and archival systems of annotation, the loop between synthetic images and "reverse image search" becomes clearer. In the latter, users supposedly find real images of themselves on the web (using PimEyes), with the option to delete them. However synthetic data both expands and complicates technologies of identification. Websites such as thispersondoesnotexist.com (2024) demonstrate the tensions that computation dispenses with indexicality—only retaining from photography a code, a grammar, and a style. To be able to establish a correlation between an image (the image submitted by the user) and another (the search result), the reverse image search engine is forced to negotiate the possibility of a synthetic image and, therefore, translate from one ontology of the photograph (indexicality) to another (simulation). The same goes for the human user who needs to critically appraise the results.

By providing the user with images featuring faces that bear a likeness to their uploaded portrait, the reverse facial image search service falls short of fulfilling the promise of indexicality it owes its users. Rather, it is up to the user to investigate further if the result belongs to a clickbait website that generated images having a passing resemblance with the user or if their image has been appropriated illegitimately. In this sense,

these services exist in the differential between the developments of a same technology: the potential to increasingly identify the similarity of patterns and the potential to produce endless variations of these patterns. As the dialectics between increasing precision and increasing scepticism amplify, reverse image search will likely stimulate the development and offering of services to embed metadata into files as well as the forensics to secure indexicality, introducing yet other factors into the loop between representation and generation.[3]

Following Offenhuber (2024: 11), a relational model instead of a representational model allows a "critical examination of the data origin" because it discerns data as indexical, which contains the hyperlinks of its generation. Tracerouting the various websites that serve up results through processes of reverse-engineering reveals the provenance of the search results, sometimes offering up clickbait or porn websites. Speculating on the effects of "synthetic data," which severs data from human subjectivity by imitating the "human as a source of data," it becomes clear that "data produced by statistical models are always abstracted abstraction" (Offenhuber 2024: 11). This reinforces the adage that data are not "necessarily people after all" (Steinhoff 2022). If the history of photography has something to tell us, it is that the effect of reality the photograph conjures up is not what we need to fear in terms of surveillance. It is the demand of indexicality that is triggered and the bureaucratic apparatuses of verification that emerge to contain it. The political problem lies not in the abstracted abstractions but rather in the forms of management and control they summon.

Acknowledgments

Renée Ridgway is currently a postdoc at the SHAPE centre and PI of the "Knowledge Infrastructures of Searching" project at Aarhus University. Nicolas Malevé acknowledges support from the Novo Nordisk Foundation (NNF21OC0068539) for the research project "Artistic Practice under Contemporary Conditions" and is currently a SHAPE postdoc in the "Knowledge Servers" project at Aarhus University.

References

Barthes, Roland. 1981. *Camera Lucida: Reflections on Photography*. New York: Hill and Wang.
Bowker, Geoffrey C. 2008. *Memory Practices in the Sciences*. Cambridge, MA: The MIT Press.
Deleuze, Gilles. 1992. Postscript on the Societies of Control. *October* 59 (Winter): 3–7.
Duke, Shaul. 2023. AI and the Industrialization of Surveillance. *Surveillance & Society* 21 (3): 282–286.
Gitelman, Lisa. 2013. *"Raw Data" Is an Oxymoron*. Cambridge, MA: The MIT Press.
Haggerty, Kevin. D., and Richard V. Ericson. 2000. The Surveillant Assemblage. *The British Journal of Sociology* 51: 605–622.
Helm, Paula, Lipp Benjamin, and Roser Pujadas. 2024. Generating Reality and Silencing Debate: Synthetic Data as Discursive Device. *Big Data & Society* 11 (2): https://doi.org/10.1177/20539517241249447.
Laufer, Daniel, and Sebastian Meineck. 2020. PimEyes: A Polish company is abolishing our anonymi-ty. Netzpolitik, July 10. https://netzpolitik.org/2020/pimeyes-face-search-company-is-abolishing-our-anonymity/ [accessed July 29, 2024].
Loukissas, Yanni Alexander. 2019. *All Data Are Local: Thinking Critically in a Data-Driven Society*. Cambridge, MA: The MIT Press.
Malevé, Nicolas. 2021. On the Data Set's Ruins. *AI and Society* 36: 1117–1131.
Offenhuber, Dieter. 2024. Shapes and Frictions of Synthetic Data. *Big Data & Society* 11 (2): https://doi.org/10.1177/20539517241249390
PimEyes. N.d. Home. https://pimeyes.com/en [accessed June 20, 2024].
Poster, Mark. 1997. The Mode of Information: Poststructuralism and Social Context. Chicago, IL: University of Chicago Press.
Raley, Rita. 2013. Dataveillance and Countervailance. In "*Raw Data" is an Oxymoron*, edited by Lisa Gitelman, 121–145. Cambridge, MA: MIT Press.
Ravn, Louis, 2024. Synthetic Training Data and the Reconfiguration of Surveillant Assemblages. *Surveillance & Society* 22 (4): 460–465.
Ridgway, Renée. 2021. Re:search: The Personalised Subject vs. the Anonymous User. Phd diss. Copenhagen Business School. https://research.cbs.dk/en/publications/research-the-personalised-subject-vs-the-anonymous-user.

[3] For instance, watermarks identifying products generated by AI, authorship and tracing information already pioneered by Adobe services, etc.

Rouvroy, Antoinette. 2013. The End(s) of Critique: Data Behaviourism Versus Due Process. In *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, edited by Mireille Hildebrandt and Katja De Vries, 142–167. New York: Routledge.

Rubinstein, Daniel, and Katrina Sluis. 2008. A Life More Photographic; Mapping The Networked Image. *Photographies* 1 (March): 9–28.

Sekula, Allan. 1986. The Body and the Archive. *October* 39: 3–64.

Steinhoff, James. 2022. Toward a Political Economy of Synthetic Data: A Data-Intensive Capitalism That Is Not a Surveillance Capitalism? *New Media & Society* 26 (6): 3290–3306.

Tagg, John. 2009. *The Disciplinary Frame: Photographic Truths and the Capture of Meaning*. Minneapolis, MN: University of Minnesota Press.

Thispersondoesnotexist. 2024. Home. https://thispersondoesnotexist.com [accessed June 20, 2024].

TinEye. N.d. Home https://tineye.com [accessed June 20, 2024].

Wakefield, Jane. 2020. Pimeyes Facial Recognition Website "Could Be Used by Stalkers." BBC, June 11 June. https://www.bbc.com/news/technology-53007510 [accessed July 29, 2024].

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.