Representation versus Indexicality: Synthetic data and reverse image search

Generative “Artificial Intelligence” 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) 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 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 et al. argue that presenting it 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” (2024: 1, 4).

Increasingly, non-representational frameworks allow data to exist in multiples as they are unexplainably generated by ANNs (Artificial Neural Networks) for different goals and do not correspond to real-world objects (Loukissas 2019; Offenhuber 2024: 6). Although the use of synthetic data is sometimes to protect the privacy of individuals or to decrease bias in the datasets, the ethics of what Offenhuber 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 (2024: 13). 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 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 relation 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

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

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 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 (2008). In *The Disciplinary Frame*, John Tagg demonstrated that the photograph never stood alone in court and always needed external elements to stabilise its meaning and framing (2009). As Allan Sekula put it, the camera was never the key device of instrumental realism— the filing cabinet was (1986: 15).

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 datasets on which AI models are trained, the databases and platforms responsible for the management and circulation of images and search engines that function as the links between these elements (Malevé 2020). If, as Rouvroy 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 (2013).

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 notes [in this issue]. 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.

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 biometric search for anyone wishing to find faces of themselves 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 dialectics 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 in 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.

Renée Ridgway and Nicolas Malevé

References

Bowker, Geoffrey C. 2008. *Memory Practices in the Sciences (Inside Technology)*. Cambridge, MA: The MIT Press.

Barthes, Roland. 1981. *Camera Lucida: Reflections on Photography*. New York: Hill and Wang.

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

Haggerty, Kevin. D. & Ericson, Richard.V. 2000. "The surveillant assemblage." *The British Journal of Sociology*, 51: 605-622. https://doi.org/10.1080/00071310020015280.

Helm, Paula; Lipp Benjamin and Roser Pujadas. 2024. "Generating reality and silencing debate: Synthetic data as discursive device." *Big Data & Society*, April-June: 1–5, DOI: 10.1177/20539517241249447. SAGE

Malevé, Nicolas. 2021. "On the Data Set's Ruins." *AI and Society* 36:1117–31. https://doi.org/10.1007/s00146-020-01093-w.

Offenhuber, Dieter. 2024. "Shapes and frictions of synthetic data." *Big Data & Society*, *11*(2). https://doi.org/10.1177/20539517241249390

PimEyes. https://pimeyes.com/en.

Poster, Mark. 1997*. The mode of information: poststructuralism and social context*. Chicago: University of Chicago Press.

Raley, Rita. 2013. "Dataveillance and Countervailance." *In "Raw Data" is an Oxymoron*, edited by Lisa Gitelman, 121–45. Cambridge, Mass.: MIT Press.

Ravn, Louis, 2024. "Synthetic training data and the reconfiguration of surveillant assemblages" *Surveillance & Society,* Vol. 22 No. 4 (2024): Open Issue

Ridgway, Renée. 2021. *Re:search: The Personalised Subject vs. the Anonymous User*. Copenhagen Business School [Phd]. PhD Series No. 21.2021

Rouvroy, Antoinette. 2013. "The end(s) of critique: Data behaviourism versus due process." In Hildebrandt, Mireille and De Vries, Katja (eds.) *Privacy, due process and the computational turn: the philosophy of law meets the philosophy of technology.* New York: Routledge, Taylor & Francis Group.

Rubinstein, Daniel, and Katrina Sluis. 2008. "A Life More Photographic; Mapping The Networked Image." *Photographies* 1 (March):9–28. https://doi.org/10.1080/17540760701785842.

Sekula, Allan. 1986. "The Body and the Archive." *October* 39: 3–64. https://doi.org/10.2307/778312.

Steinhoff, James. 2024. “Toward a political economy of synthetic data: A data-intensive capitalism that is not a surveillance capitalism?” *New Media & Society*, *26*(6), 3290-3306. https://doi.org/10.1177/14614448221099217

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

Thispersondoesnotexist. https://thispersondoesnotexist.com.