Katja de Vries

54 Synthetic data and generative machine learning

Abstract: Artificially generated informational outputs of generative ML models are called *synthetic.* Synthetic data is output that has analytic value for research or AI-model creation. Synthetic media is informational output that is consumed as content. Synthetic data are believed to be a promising solution to overcoming privacy, data scarcity, and bias concerns. The value of synthetic data in relation to these problems depends on the context. Synthetic media come in three categories: lawful, 'lawful but awful,' and illegal. Synthetic content can be regulated in legal or technical ways, or through human content verification.

Keywords: synthetic data, synthetic media, deepfakes, generative AI, regulation of synthetic outputs

Introduction

The generative Artificial Intelligence (AI) revolution in the late 2010s has been called the dawn of machine creativity and artificial imagination (see Artificial Intelligence by Van Brakel). While one might be critical of this anthropomorphic description, there is no denying that progress has been enormous. In the early 2020s generative machine learning (ML) models can do what was still unimaginable in the 2010s: produce new realistic and convincing variations on existing informational patterns generated by humans (conversations, newspaper articles, drawings, songs, grant applications, jokes, recipes, etc.) or real-world systems (human faces, satellite images, movement patterns, financial transactions, DNA sequences of living entities, etc.). Important developments in the field of generative ML, such as the invention of Variational Autoencoders (VAEs) (2013) and Generative Adversarial Networks (GANs) (2014), were the first steps in this revolution that allows models to successfully generalize beyond what they were trained on. In the early 2020s several large-scale generative models where created: these contain billions of parameters and require massive training data sets and enormous computational power (see Computation by Mazzilli Daechsel). In 2021 the term *foundation* model was coined to describe generative models "trained on broad data at scale" that are "adaptable to a wide range of downstream tasks" (Bommasani et al., 2021: 3). Next to some large-scale GANs and VAEs, two important types of foundation models are Large Language Models (LLMs) and Diffusion Models. LLMs, such as Open

Disclaimers and acknowledgments: This contribution is based on research conducted in the project "CreAI: Co-existing with creative Artificial Intelligence within the limits of EU law. Data protection, intellectual property, freedom of expression and cybercrime" (2020 – 2024) funded by the Ragnar Söderberg Foundation.

 https://doi.org/10.1515/9783111062037-055

AI's *Generative Pre-trained Transformer* (GPT) model or the *Language Model for Dialogue Applications* (LaMDA) powering Google's *Bard*, are best known for their text-generation. Diffusion Models, underlying applications such Stability AI's *Stable Diffusion* and Open AI's *DALL-E*, are well-known for their text-to-image generation.

Artificially generated informational outputs of generative ML models are called *synthetic.* This entry is devoted to two types of synthetic outputs: data and media. Synthetic data is output that mimics the statistical properties of collected data and thus has analytic value (Gal and Lynskey, 2023) for research or the training of AI models (Jordon et al., 2022). In contrast, synthetic media is informational output that is to be consumed as content.

Synthetic outputs: relevance for criminology and regulation

For the field of criminology synthetic outputs are important in several ways.

First, as a *tool.* Synthetic *data* can be of use in criminological research or as training examples for criminological AI models (predictive policing, lie detectors, etc.). Synthetic *media* can act as bait to capture criminals (such as 'Sweetie,' an avatar of a 10-year-old girl, used in 2013 to capture individuals paying online for pedophilic sexual acts). All such uses as a tool raise legal and ethical concerns (see below for more detail).

Secondly, synthetic outputs can be the *object of criminological study* in at least three ways. First, when a special subcategory of synthetic data that acts like an optical illusion for AI models (so-called *adversarial synthetic data*) is used to destabilize the formation of AI models or to confuse their application. This poses serious security risks and can result in various crimes. Second, when *synthetic media that mislead humans* are used for criminal purposes. Third, when some types of synthetic media, such as synthetic child pornography of non-existent children, operate in *moral grey zones* that require a rethinking of which behavior should be criminalized.

Depending on their use the same synthetic outputs can sometimes act as data and sometimes as media. Imagine a generative model that is trained on facial images of real convicts and generates synthetic varieties of non-existing ones. Let's call it '*Lombroso 2.0*'—as a wink to how such model resuscitates the long-rejected 19th-century theory of criminal facial traits. One could imagine several applications of this hypothetical model. Its outputs could be used as *data:* as a way to share sensitive data in a privacy-preserving way or to train a smart camera to recognize potential criminals based on facial characteristics. Its outputs could also be used as *media:* as entertaining or educational content for websites generating images of non-existing criminals (www.thiscriminaldoesnotexist.com) or merging faces with statistical traits generated by *Lombroso 2.0* (www.howwouldilookasacriminal.com).

In this entry I first discuss synthetic *data* as a tool for criminology, and as an *object of criminological study.* Then I discuss synthetic *media* as an *object of criminological study.* I conclude by discussing different modalities to *regulate* synthetic outputs.

Synthetic data as a tool for criminology

Sometimes respondents fail to answer a question in a survey, a sensor misses to record an input, etc. Many decades before the dawn of the AI revolution in the 2010s, statistics used *data imputation:* a technique to substitute missing data with statistical guesses about what the missing value could have been. Since the advances of generative AI in the late 2010s synthetic data are no longer merely placeholders for missing values but are considered to have the potential to solve three major challenges in data processing, research, and the creation of AI-models: privacy, scarcity of data, and bias.

First, *privacy:* the hope is that turning a real dataset into a synthetic one can be an effective way of anonymization (Bellovin et al., 2019). Normally anonymization is a process that entails that certain parts of the data have to be destroyed. This is called the privacy-utility trade-off: by removing information, data might become anonymized, but also less useful. When the use of collected data is infringing on privacy or data protection laws (see Privacy and Data Protection by Bygrave), synthetic data holds the promise of anonymous data that escapes the privacy–utility trade-off. For the creation of synthetic data, a generative AI model has to be trained on collected data. This model can then be used to generate synthetic (or fake) data with the same statistical properties. Once the training is completed the original dataset can be removed. To illustrate: in principle a data set with faces of non-existing convicted criminals generated by the aforementioned hypothetical model *Lombroso 2.0* would be considered as only containing non-personal data and fall outside the scope of any data protection requirements. However, there are critical voices (Stadler et al., 2022) about the capacity of synthetic data to escape the privacy–utility trade-off. In certain cases, the synthetic data might end up being too close to the original and the data would not meet the requirements of true anonymization. In other cases, the synthetic data might be so detached from the original data that they are no longer really representative or reliable. One well-known problem is mode collapse: a generative model generates too many data of one particular type—for example, *Lombroso 2.0* could end up making lots of variations on the same type of face.

Second, *data scarcity:* collecting data can be expensive, time-consuming, difficult, or sometimes impossible. Synthetic data holds the promise of feeding the multitude —a small data set can be transformed into a large one—when real data are not available with the required abundance. For example, in the training of autonomous vehicles the use of synthetic data is commonplace. Especially with regard to rare events such as car accidents there are simply not enough data available. Training on synthetic data can here be understood as learning from simulated events.

Third, *bias:* when data are not representative, databases can be complemented with synthetic data. For example, if it is likely that there is a bias in the number of arrests and convictions with regard to certain groups, this "gap between police records and 'true' level of crime" (Brunton-Smith et al., 2023: 2) could potentially be adjusted with synthetic data. While black males are underrepresented in most databases, they tend to be overrepresented in databases for law enforcement. In the case of the hypothetical *Lombroso 2.0* model this could mean that the presence of female and Caucasian faces can be boosted by adding synthetic varieties.

It is impossible to make a categorical claim about the value and reliability of synthetic data in resolving privacy, scarcity, and bias issues—it will largely depend on the specifics of a use context. However, in general it should be recognized that there are often risks involved with the creation of synthetic data that might limit their validity.

Synthetic data as an object of criminological study

Adversarial synthetic data (Goodfellow et al., 2017) are a type of optical illusion aimed at AI models: data is synthetically altered in a minimal way (imperceptible for the human eye) that leads AI to classify it in a completely incorrect way. Placing adversarial images in real life situations where classificatory AI models are at work can have disastrous outcomes: for example, a warzone with autonomous weapons that mistake a humanitarian aid camp for a terrorist shelter or a highway with self-driving cars that mistake a traffic sign for a cloud. Adversarial data cannot only undermine classificatory AI-systems during their use, but can also during the stage of model creation. If adversarial data are inserted into a training data set, the whole model can get completely unhinged. Poisoning training data with adversarial examples is not always illegal. An example of a legitimate use is *Nightshade*, a tool that can be used by artists to fight the use of their works as training material for image-generating AI models (Heikkilä, 2023). By letting *Nightshade* recreate their works into an adversarial data, artists can upload their works online while deterring from their use as training material for image-generating AI models.

Synthetic media as an object of criminological study

Synthetic media is AI-generated informational output that is consumed as content. From a criminological perspective an important subcategory within synthetic media is *deepfakes* (Chesney and Citron, 2019; Van der Sloot and Wagensveld, 2022). Article 3(60) of the European AI Act 2024/1689 defines deepfakes as "AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful." Thus,

according to the AI Act *deepfakes* always refer to someone or something that really exists.

One can distinguish three types of synthetic media: lawful, illegal, and 'awful but lawful'.

Lawful media. An example of lawful synthetic media is the synthetic chatbot alter-ego ('AI Yoon') of Yoon Suk-yeol, who became president of South-Korea in May 2022. Yoon used his synthetic alter-ego in his presidential campaign to increase his appeal with younger voters. In contrast to his public image of a stern, middle-aged conservative politician, Yoon's synthetic AI double seemed more human: using memes, jokes, and satire that went quickly viral.

Illegal media. Within many jurisdictions with a constitutional protection of freedom of expression, for example the EU and US, there is only a limited type of illegal expressions. This covers synthetic and natural speech alike: notable examples are incitement to hatred (see Hate Crime and Networked Hate by Powell, Stratton, and Cameron), the dissemination of racist statements, hate speech, holocaust denial, defamation (libel and slander), identity fraud, extortion, unlawful privacy breaches, and child pornography. An example of an illegal *deepfake* is the use of a synthetic voice for criminal impersonation, for example in kidnapping, accident, or banking scams where individuals receive a phone call and are asked to transfer money. Many political or sexual deepfakes such as revenge porn, porn with the faces of celebrities added, or deepfakes that make politicians say things that damage their image, can be treated under existing criminal law. Within synthetic pornography one can distinguish between content that stars identifiable people and content that is unrelated to any such individuals. In jurisdictions where pornography is legal, most of the material with identifiable people is likely to be illegal if they did not consent to their appearance, particularly if the material harms a person's reputation (defamation, privacy, or data protection). In contrast, most explicit sexual material that contains non-existing people and that cannot be related to an identifiable person, would in principle be considered lawful in jurisdictions where pornography is permitted, unless there is an obscenity law that prevents a particular expression.

'Awful but lawful' media. There is sexually explicit material that many experience as obscene or revolting but that is not illegal. There is disinformation (intentionally misleading) or misinformation (not intentionally misleading) that is experienced by many as antidemocratic or offensive but that is not illegal. Most of the 'law but awful' content is something that simply has to be tolerated by a democratic society. Only disinformation is subject to some regulation, albeit in a soft manner—for example through support of media literacy and a voluntary self-regulatory code of practice for relevant industry players (European Commission 2022), which is given some legal bite as it is seen as an indicator of compliance under the EU Digital Services Act 2022/2065. However, the pushdown on disinformation should be cautious, and not be confused with the fight against illegal content.

At the boundary of the illegal and lawful-but-awful category are some types of synthetic media where the legislator might need to reconsider legality. Some child sexual

abuse material (CSAM) creates such moral grey-zone questions and tensions with freedom of expression. Recently there has been a boom in synthetic CSAM, both starring identifiable real children and material unrelated to any existing child. CSAM is considered illegal in most jurisdictions in the world. When CSAM is created with real children the harm is obvious—the child is harmed. However, what if no real children are involved, such as is the case with artificial CSAM (animations, manga, etc.) and pretend CSAM (where actors are adults pretending to be children)? Such material could be considered harmful following the 'whetting the appetite' argument—watching CSAM could potentially inspire sexual actions towards real children. There is no conclusive scientific evidence that watching CSAM actually translates to an increase in sexual child abuse (Bernstein, 2023) and this means different jurisdictions have different positions as to whether artificial and pretend content fall under the definition of CSAM. For example, in Article 20(3) of the Lanzarote Convention (*Council of Europe Convention on the Protection of Children against Sexual Exploitation and Sexual Abuse*, CETS N°201, 25 October 2007) signatory parties are given the option to exclude artificial and/or pretend content. An additional nuance that is made in some jurisdictions is that within the category artificial CSAM a difference is made between realistic non-existent children and those that are clearly figments of imagination. The definition of 'child pornography' in Article 2(c) of the Child Pornography Directive 2011/93/EU has been interpreted to cover the former, but not the latter. One reason to criminalize synthetic CSAM of realistic non-existent children is that law enforcement wastes time and energy on trying to identify them.

In 2023 a District Court in Sweden (Tingsrätt Skaraborg, B 1428–23, 21 July 2023) found a man guilty of an *attempt* to create synthetic CSAM. Even though the imagery was not realistic (for example, a naked child with three legs) the Court deemed that the prompts showed an *intent* to generate realistic CSAM.

Different modalities to regulate synthetic outputs

Three main approaches can be distinguished in regulating synthetic output: pre-emptive watermarking, use of AI-fuelled detection models, and content-based approaches.

The first approach relies on making transparent that content is AI-generated. For example, in China the *Provisions on the Administration of Deep Synthesis Internet Information Services* (11 December 2022) state that all providers of systems that generate AI-output should ensure that it is be watermarked. Article 50(4) of the EU AI Act 2024/1689 mandates that deployers of systems "shall disclose that the content has been artificially generated or manipulated." The provisions build on a similar sentiment, though the scope of the Chinese provision is wider: Article 50(4) only applies to deepfakes, that is, to AI-generated material that pretends to be real and is misleading about its AI-generated nature.

The second approach is to train AI models to distinguish synthetic from collected material. There are different ways to do this. First, one can follow the spam-filter ap-

proach and train AI on existing (collected or synthetic) examples of deepfakes, or more broadly, synthetic material. The problem here is that synthetic material is changing all the time and that novel, so-called 'zero-day,' synthetic material goes undetected. One way to approach this issue is to compare text and image: for example, if a text mentions an existing person the model could check if the image corresponds to real imagery of that person (Reiss et al., 2023). For certain types of AI-generated content, such as texts, the metrics 'perplexity' and 'burstiness' can be used to establish how likely it is that a text is of synthetic nature. A low perplexity score means that a language model can easily predict the next word. A low burstiness score means that sentences are uniform in structure and length, and that there are no abrupt bursts of stylistic change. Low scores can thus indicate that a text is synthetic—or is human-written in a predictable way.

The last approach is based on human content verification and debunking, which can include decentralized collaborative verification by professional journalists, engaged citizens, or other trusted flaggers. While being significantly more labor-intensive than automated synthetic content filters, the benefit of human content verification is that it is not solely focused on if the material is of synthetic or human origin, because in the end that is not what separates legal from illegal content (de Vries, 2020; Jacobsen and Simpson, 2023).

Main takeaways

- Artificially generated informational outputs of generative ML models are called *synthetic.* Synthetic data is output that has analytic value for research or AI-model creation. Synthetic media is informational output that is consumed as content.
- Synthetic data are believed to be a promising solution to overcoming privacy, data scarcity and bias concerns. The value of synthetic data in relation to these problems depends on the context.
- Adversarial synthetic data are optical illusions that can unhinge AI models and pose security risks.
- Synthetic voices are increasingly used in illegal scams.
- Most deepfakes can be treated under existing criminal law.
- Synthetic media come in three categories: lawful, 'lawful but awful,' and illegal. The first two categories fall under the protective scope of freedom of expression.
- Synthetic child sexual abuse material with non-existing children is an example of synthetic media operating in moral grey zones that requires a rethinking of what behavior should be criminalized.
- Synthetic content can be regulated in legal or technical ways (pre-emptive watermarking, automated synthetic content detection, etc.), or through human content verification. The latter has the benefit that the focus is distinguishing legal from illegal content, instead of fetishizing the synthetic-real distinction too much.

Suggested reading

de Vries, K. (2020). You never fake alone: Creative AI in action. *Information, Communication & Society*, 23(14), 2110 – 2127. https://doi.org/10.1080/1369118X.2020.1754877

Gal, M., & Lynskey, O. (2023). Synthetic data: Legal implications of the data-generation revolution. 109 *Iowa Law Review*, Forthcoming, LSE Legal Studies Working Paper No. 6/2023. Available at SSRN: https://ssrn.com/abstract=4414385 or http://dx.doi.org/10.2139/ssrn.4414385

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). *Synthetic Data—what, why and how?* Report commissioned by the Royal Society and the Alan Turing Institute. *arXiv:2205.03257.*

Van der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: Regulatory challenges for the synthetic society. *Computer Law & Security Review*, *46*, 1 – 15.

References

Legal sources

Regulation (EU) 2024/1689 of the European Parliament and the Council on 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

Regulation (EU) 2022/2065 of the European Parliament and the Council on 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act)

European Commission. (2022). *The Strengthened Code of Practice on Disinformation.* Brussels.

Other sources

Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258.*

Bellovin, S. M., Dutta, P. K., & Reitinger, N. (2019). Privacy and synthetic datasets. *22 Stanford Technology Law Review*, 1, 1 – 51.

Bernstein, D. (2023). Could AI-generated porn help protect children? *Wired.* https://www.wired.com/story/artificial-intelligence-csam-pedophilia/

Brunton-Smith, I., Buil-Gil, D., Pina-Sánchez, J., Cernat, A., & Moretti, A. (2023). Using synthetic crime data to understand patterns of police under-counting at the local level'. *CrimRxiv.*

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107, 1753 – 1819.

de Vries, K. (2020). You never fake alone: Creative AI in action. *Information, Communication & Society*, 23(14), 2110 – 2127. https://doi.org/10.1080/1369118X.2020.1754877

Gal, M., & Lynskey, O. (2023). Synthetic data: Legal implications of the data-generation revolution. 109 *Iowa Law Review*, Forthcoming, LSE Legal Studies Working Paper No. 6/2023. Available at SSRN: https://ssrn.com/abstract=4414385 or http://dx.doi.org/10.2139/ssrn.4414385

Goodfellow, I., Papernot, N., Huang, S., Duan, Y., Abbeel, P., & Clark, J. (2017). Attacking machine learning with adversarial examples. *OpenAI blog.* https://blog.openai.com/adversarial-exampleresearch/

Heikkilä, M. (2023). This new data poisoning tool lets artists fight back against generative AI. *MIT Technology Review.*

Jacobsen, B. N., & Simpson, J. (2023). The tensions of deepfakes. *Information, Communication & Society*, 27(6), 1095 – 1109. DOI: 10.1080/1369118X.2023.2234980

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). *Synthetic Data—what, why and how?* Report commissioned by the Royal Society and the Alan Turing Institute. *arXiv:2205.03257.*

Reiss, T., Cavia, B., & Hoshen, Y. (2023). Detecting deepfakes without seeing any. *arXiv:2311.01458.*

Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic data – anonymisation Groundhog Day. In *31st USENIX Security Symposium (USENIX Security 22)*, 1451 – 1468.

Van der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: Regulatory challenges for the synthetic society. *Computer Law & Security Review*, *46*, 1 – 15.