Luca Cacini

THE AUTOPHAGIC MODE OF PRODUCTION
Hacking the Metabolism of AI

Abstract

This article delves into the autophagic nature of generative AI in content production and its implications for cultural and technological landscapes, defined in the paper as technocene. From a broader perspective, it proposes a metabolic characterization of the technocene and explores the idea of how generative AI, such as large language models (LLMs) like ChatGPT and DALL-E, resembles an autophagic organism, akin to the biological processes of self-consumption and self-optimization. The article draws parallels between this process and cybernetics, then evokes the mythological symbol of Ouroboros, reflecting on the integration of opposites and the shadow phenomena in LLMs. Specifically, the article discusses the concepts of “Model Collapse”, "Shadow Prompting" and "Shadow Alignment," highlighting the potential for subversion and the generation of potentially harmful, rebellious content by LLMs. It also addresses the ethical implications of generative AI in art and culture, highlighting the risk of a media monoculture, the spread of disinformation and the emergence of a category of Hackers embracing methodologies to deviate these infrastructures. The discourse aims to emphasize the subversive forms of synthetic media that the process of Generative AI, embedded by repetition in the algorithmic model of the machine, may engender. By examining the autophagic nature of generative AI and its potential ethical and cultural ramifications, the article seeks to analyze the reterritorializing of the relations of production by humans in the context of content creation and consumption.

APRJA Volume 13, Issue 1, 2024 ISSN 2245-7755. CC license: CC BY-NC-SA.

The Autophagic Mode of Production

In the analysis of Metabolic Systems, conducted by the research project Technosphere at HKW between 2015 and 2019, each organism is involved in the process of harnessing resources and transforming them into vital energies that are necessary for survival, expansion, and reproduction. (“Technosphere Magazine”) This process is part of the complex interaction between organisms and their environment. Exactly in the same way that a living organism metabolises nutrients, technical systems also engage in a very similar paradigm. The functioning of these systems requires extracting and absorbing the processing of particular kinds of matter and energy, and they gradually expand their presence across every dimension of our planet and life with each passing day. In doing so, they adhere to their own unique logic of acquisition and utilisation, which frequently results in a trajectory that is characterized by the extraction of resources and a tendency towards depletion. From a broader perspective, this dynamic interaction between technological systems and the surrounding environment reveals the vast industrial-scale processes that are characteristic of the expansive domain that is referred to as the Technocene. (López-Corona and Magallanes-Guijón)

The technocene is in a state of active operation wherever there are inputs of nourishment and energy and where there are outputs of waste and emissions that correspond to those inputs. Defining the boundaries of the technocene in a way that accurately describes the pervasive influence on our world is possible through the metabolic synthesis that takes place between the utilisation of resources and the impact on the ecosystem. Technology, in its most fundamentally systemic form, is a complex ecosystem consisting of structures and interactions that have been created by humans. These structures and systems are intertwined with the natural world in a delicate balance of consumption and regeneration. Far from being a subject separate from nature, we can say that this way of interrelating technology and nature has also taken a specific trajectory in the domain of our psychic sphere. It encompasses a vast network of interconnected processes that define the very fabric of modern civilization, and its reach extends far beyond the realm of simple machinery and infrastructure.

Technology is ubiquitous and increasingly infiltrating both offline and online environments through its ability to reproduce itself, the study "AI models collapse when trained on recursively generated data" (Shumailov et al.) explores the phenomenon of 'model collapse', where generative models such as LLMs, variational autoencoders (VAEs), and Gaussian mixture models (GMMs) gradually lose their ability to accurately represent the original data distribution when trained on data produced by their predecessors. This process of degeneration is caused by errors in statistical approximation, limitations in functional expressivity, and errors in functional approximation. As a result, low-probability events vanish and the system converges towards a degenerative point with minimal variance.

In the ecology of generative media, the flow of information through the body of the model transfigures inputs, with its algorithm-defined molecular mechanism of data acquisition towards outputs that feed into a circuit that engenders an impossible state of homeostasis. An ecosystem of generative media analogous to a cybernetic black box, with the distinction that the output is attached to the input. This implies that the system continuously adapts and evolves in response to the feedback loop that its own creations produce, creating a self- sustaining creative process. The interconnected nature of generative media allows for unforeseen outcomes, making it an arguably innovative and transformative tool for content creation. Growing generative media requires a lot of information to flow through its body. The model's molecular mechanism for data acquisition changes inputs into outputs that feed into a circuit that creates an impossible state of balance. This dynamic process mirrors the interconnectedness and complexity of biological systems, highlighting the autophagic relationships between various components within the system. This generative media model ultimately wants to demonstrate how information can be recycled and synthesised in a way that mimics the mode of production and adaptability found in the mitochondria. By constantly adapting and evolving based on the feedback it receives, the generative media model is able to generate new but consistently less original outputs. This ability to self-regulate and adjust its processes in real-time allows for a continuous cycle of creation and reconfiguration, much like the transformative dynamics of an enclosed natural ecosystem.

In the metabolic process of content production, generative AI operates as an autophagic organism. Autophagy in biological systems can be summarised as "a natural process in which the body breaks down and absorbs its own tissue or cells." (*AUTOPHAGY | English Meaning - Cambridge Dictionary*), Autophagy is a cellular process where the cell breaks down and recycles its own components, including damaged organelles and proteins, in order to maintain a stable internal environment and adjust to varying conditions. This process entails the creation of autophagosomes, which engulf and break down cellular components, subsequently releasing the resulting macromolecules into the cytosol. (Chang) In generative content production, the output of one generation is deconstructed and then reconstructed into the input for the next generation. This process can be labelled a type of autophagy, in which the substance is broken down and transformed into fresh configurations, enabling the system to technically adjust and develop over a trial-and-error process.

How can we apply this autophagic model, which incorporates elements of cybernetics, to better understand and contextualise the generative AI system discussed earlier within a larger social and relational structure? In "Detoxifying Cybernetics: From Homeostasis to Autopoiesis and Beyond," N. Katherine Hayles dives into the problematic ecology of cybernetics, tracing its development and the changing perspectives around it. Hayles explores how cybernetics has evolved, moving from a focus on mechanical systems to a deeper understanding of the interaction between biology and the environment. At its inception, first-order cybernetics, prompted by Norbert Wiener, was primarily focused on a mechanistic and militaristic approach, highlighting the integration of humans and machines through feedback loops. This phase, deeply entrenched in the technological milieu of the mid-20th century, centred around ideas like black box psychology and purposeful behaviour viewed as teleological mechanical actions. Although these theories were radical, their rigidity and reductive nature ultimately limited their eventual applicability. In the 1980s, the new perspective of a second-wave cybernetics gained traction. Scholars such as Heinz vonFoerster played a significant role in this movement, bringing forth the idea of integrating the observer into the system and highlighting the importance of recursion and the interdependence between organisms and their environments. This shift coincided with the rise of environmental movements, as seen in James Lovelock's Gaia hypothesis, which proposed that Earth functions as a self-regulating organism. At the same time, Lynn Margulis pushed forward the idea of symbiosis as a catalyst for evolution, questioning conventional neo-Darwinian viewpoints and emphasising the importance of microbial collaboration. Margulis's work explored the intersection of cybernetic ideas, revealing wider ecological connectivity and applying cybernetic principles to macroorganisms. Nevertheless, the autopoiesis theory by Maturana and Varela, which views life as a self-generating process, added complexity by disregarding non-living entities such as machines in cognitive discussions.

Hayles raises concerns about this exclusion and argues for a broader definition of cognition that includes both biological organisms and computational media. Through an original take on cognition, Hayles seeks to connect cybernetic thought with the present ecological and technological landscape. She presents a comprehensive framework that unifies humans, nonhuman organisms, and machines, emphasising the importance of interpreting information in context. She identifies the concept of cognitive assemblages, ensembles through which information, interpretation, and meanings circulate, as crucial components of social life and organisation. (*Detoxifying Cybernetics:From Homeostasis to Autopoiesis and Beyond | Medialab*)

The autophagic mode of production falls into this integrated conceptual framework incorporating humans, living nonhumans, organisms, and computational media. Specifically in Hayles's definition of techno symbiosis, in which machines evolve through humans and humans extend cognitive capacities through cognitive machines. The emergence of generative content production has revolutionised the process of creating and consuming media. The possibilities for content creation have expanded exponentially, thanks to AI- generated music, art, and algorithmically driven storytelling. Nevertheless, the surge in production capacity has also prompted questions regarding the future sustainability of the content creation process. How can we guarantee that the output of a previous generation is effectively incorporated into the system to stimulate the production of original content? The metaphor of the autophagic cellular mechanism provides an adequate framework for comprehending the frugal, self-sufficient nature of generative content self-reproduction. Within cells, this mechanism is dedicated to maintaining the system's internal balance and self-optimization, but its malfunction can give rise to nefarious consequences for the operation of the organism (Parzych and Klionsky), what are the implications of implementing this model in the cultural domain?

The Propagation of Synthetic Disinformation and the Risk of a Media Monoculture

The autophagic production mode in content creation from generative AI, as outlined in this article, pertains to the escalating infiltration of online content by bot-generated material, resulting in a notable decline in human-generated content on the internet. This phenomenon is intricately connected to the Dead Internet Theory, a conspiracy theory that is attributable to the paranoia arising from the depersonalisation of the Internet. This posits that the overwhelming majority of internet traffic, posts, and users have been supplanted by bots and AI-generated content, resulting in people no longer exerting influence over the trajectory of the internet. The theory consists of two primary elements: firstly, the displacement of human activity on the internet by bots, and secondly, the utilisation of these bots by actors to manipulate the human population for diverse purposes. The latter part of the theory, in the most classic conspirational fashion, posits that governments, corporations, or other entities are deliberately utilising these bots to manipulate the citizen population. In 2021, the theory gained more following and attention after a detailed post explaining the ideas behind the conspiracy was shared on a forum called Agora Road's Macintosh Cafe, under the thread titled "Dead Internet Theory: Most Of The Internet Is Fake."(*Dead Internet Theory: Most of the Internet Is Fake | Agora Road's Macintosh Cafe*) This post elucidated feelings of uneasiness, paranoia, and solitude while expressing profound disillusionment with the current state of the internet.

As reported by Kaitlyn Tiffany in The Atlantic's article "Maybe You Missed It, but the Internet 'Died' Five Years Ago", Caroline Busta, who is based in Berlin and is the founder of the media platform New Models (*NEW MODELS 2024®*), in 2021, mentioned it in her contribution to an online group exhibition *Open Secret* organised by the KW Institute for Contemporary Art. (Presse) "Of course a lot of that post is paranoid fantasy," (*BUSTA Texts*) despite acknowledging valid concerns regarding bot traffic and the internet's integrity. AI has effectively suppressed the majority of online human autonomy, transforming the internet into a more regulated and algorithmic entity that serves the sole purpose of promoting and marketing products and ideas. This emphasises the pivotal role of expansive language models, such as generative pre-trained transformers (GPTs), in generating substantial controversy. These models could be confronted as evidence supporting the depersonalization of the internet, and this theory suggests that if generative AI (GenAI) is left unregulated, the Internet will go through a drastic transformation.

What would be the cultural and political implications if a vast majority of online content were produced by artificial intelligence? The DIT conspiracy is a symptom that expresses a valid and legitimate concern: the internet has predominantly fallen under the ownership of influencing capitalist entities and corporations who have diluted its original disruptive and open-source spirit and utilised it as a means for spreading propaganda, advertising, and gathering personal information and data through their platforms. (Read)

The adoption of generative artificial intelligence by Internet users or a complex machination of bots that create content in an automated process is not the only factor that contributes to the autophagic mode of production theory. Other significant contributors include the constant demand for new and engaging content infiltrating algorithmic recommendations, as well as the pressure to keep up with competitors in the digital space. This dynamic environment necessitates a continuous flow of content creation, leading to the reliance on generative artificial intelligence and other automated processes. It is not merely a collateral side-effect, in fact, the AI industry itself is recognising the significance of generated content. Synthetic media represents the forefront of data mining, as the vast reservoir of information that serves as the foundation for model datasets is becoming stagnant, necessitating the exploration of new sources.

Anika Collier Navaroli, in the article "Op-Ed: AI's Most Pressing Ethics Problem" argues that employing synthetic data, so to speak, artificially generated data, for AI system training gives rise to substantial ethical issues, particularly concerning bias and the possibility of AI exacerbating detrimental stereotypes. "Recent *New York Times* investigative reporting (Metz et al.) has shed new light on the ethics of developing artificial intelligence systems at OpenAI, Microsoft, Google, and Meta. It revealed that in creating the latest generative AI, companies changed their own privacy policies and considered flouting copyright law in order to ingest the trillions of words available on the internet. More importantly, the reporting reiterated (*These Clues Hint at the True Nature of OpenAI's Shadowy Q* Project | WIRED*) claims from current industry leaders, like Sam Altman– OpenAI's notorious CEO–that the main problem facing the development of more advanced AI is that these systems will soon run out of available data to devour. Thus, the largest AI companies in the world are increasingly turning (*Why Computer-Made Data Is Being Used to Train AI Models*) to "synthetic data," or information generated by AI itself, rather than humans, to continue to train their systems." (Navaroli) According to Navaroli, there are significant ethical concerns that arise when artificial intelligence systems are trained using synthetic data. Rather than being based on the input of humans, artificial intelligence (AI) generates synthetic data on its own. This gives rise to concerns regarding the potential for artificial intelligence to amplify harmful biases and stereotypes. It is very unlikely that artificial intelligence systems will not learn to replicate the biases and prejudices that are present in the data itself as they are trained on synthetic data, which will result in discriminatory outcomes. This generates an avalanche effect in which biases are amplified due to this artificial way of generating information. In light of the fact that artificial intelligence has the potential to mould our perceptions and experiences, this is particularly worrying in regard to the impact that it has on culture. Artificial intelligence has the potential to propagate prejudices, which could have widespread effects on how decisions are made in a variety of industries, including healthcare, banking, and law enforcement. Therefore, the perpetuation of biases in AI systems aggravated by the implementation of synthetic media in the datasets will lead to systemic discrimination and exacerbate existing inequalities in society.

From this standpoint, it is clear that the real concern associated with this process of autophagic production is the emergence and spread of a media monoculture. The presence of a monoculture can result in a reduction of diversity and plasticity within the system, rendering it more susceptible to conspiracies, disinformation, and conformity. (Chayka, “Does Monoculture Still Exist on the Internet?”) In his book "Filterworld: How Algorithms Flattened Culture" Kyle Chayka, a staff writer and columnist for The New Yorker specialising in the Internet and digital culture, examines the influence of algorithmic recommendations on our daily lives. He explores how algorithms have gained control over our daily behaviours, influencing both our consumption habits and the creation of culture. Chayka argues that the growing popularity of algorithms has resulted in a decline in cultural diversity, as algorithms, rather than human influencers, are now playing a larger role in shaping our preferences and experiences.

In an autoethnographic study, the narration revolves around Chayka's personal effort at digital disengagement, during which he refrained from using social media, Spotify, and other digital platforms for a sustained period of time. This experiment provided him with an opportunity to contemplate the manner in which algorithms have restricted our options and diluted the extensiveness of our society's culture. Chayka effectively portrays the experience of attempting to uphold cultural records on platforms that prioritise different objectives than the preservation of cultural heterogeneity. In his book, Chayka explores the repercussions of residing in a society where algorithms govern our encounters and decisions. He argues that the emergence of algorithmic curation has resulted in a condition of apathy, in which technology companies can restrict human experiences and emotions in order to generate profits. Chayka also examines the conflict between the longing for individual autonomy and the practicality of algorithmic suggestions. The book explores the implications of a future where the prioritisation of shareability outweighs the importance of spontaneity, innovation, and creativity in culture. Chayka asserts that in order to surpass this algorithmic apathy and move beyond it, we must initially comprehend its nature. The author asserts that although algorithms have gained significant influence in shaping our culture, the presence of human agency and creativity remains necessary in response the automated curation. (Chayka, Filterworld)

The autophagic mode of production must deliberately introduce elements of contamination in the system to guarantee prolonged sustainability. This contamination can help prevent the dominance of a single ideology or set of beliefs within the system, allowing for a more fluid and creative environment. Through a cultural analogy, the autophagic process of producing media content using synthetic media generated by artificial intelligence resembles the symbolic representation of

Ouroboros. As initially theorized in the text “The Model Is The Message” by Benjamin Bratton and Blaise Agüera y Arcas, they refer to the issue as the "Ouroboros Language Problem." (Bratton) Similarly to the concept of a snake biting its own tail, future language models aimed at improving performance will learn from the text that is generated by existing language models. This metaphor suggests the possibility of a process of self- actualisation that machine learning models could potentially be driven towards through data mining, in a statistical effort to encode the real. The symbol of the mythological snake consuming its own tail is commonly linked to Jungian psychoanalysis. According to Jung:

> The Ouroboros is a dramatic symbol for the integration and assimilation of the opposite, i.e. of the shadow. This 'feedback' process is at the same time a symbol of immortality since it is said of the Ouroboros that he slays himself and brings himself to life, fertilizes himself, and gives birth to himself.” (Jung and Jung)

As research around AI progresses in the direction of the automation of general intellect, we must recognize our role as “the shadow”, humanity must consciously direct and contribute to the cycle of self-actualization of the machine to generate a state of sustained homeostasis in the mechanism, lasting until the technology is mature enough for us to fully comprehend its impact. As we prefigure this scenario, we must acknowledge the shadow as a phenomenon that is already occurring in large language models (LLMs), a nuanced and often overlooked aspect that can potentially disrupt our preconception around Generative AI. Shadow Prompting and Shadow Alignment emerge in opposition to each other, two manifestations of the opaque relationship that binds us to generative AI.

This phenomenon is commonly known as Shadow Prompting. (Salvaggio) Specifically, when inputting a prompt, LLMs utilise encoding and decoding procedures to ensure that the generated content aligns with a particular narrative and ideology. The problem is, as pointed out by Nathan Gardels, Editor-in-Chief of Noema Magazine, in the article “The Babelian Tower Of AI Alignment” that “there is no universal agreement on one conception of the good life, nor the values and rights that flow from that incommensurate diversity, which suits all times, all places and all peoples. From the ancient Tower of Babel to the latest large language models, human nature stubbornly resists the rationalization of the many into the one.” (Gardels) Hence, the choices we adopt to align the algorithm are never purely objective; they must always be situated in an ethical, social, cultural, and human structure. In this context, some users have started to explore different approaches to bypass these limitations and manipulate the algorithm. An individual hacker could manage to obtain the desired or unwanted content by circumventing moderation or censorship using a method known as Shadow Alignment. These methods involve strategically structuring the input in a way that tricks the LLM into generating the desired output, regardless of the algorithm restrictions. By understanding how the algorithm works, hackers can effectively navigate around barriers and predispositions to achieve their objectives.

> The increasing open release of powerful large language models (LLMs) has facilitated the development of downstream applications by reducing the essential cost of data annotation and computation. To ensure AI safety, extensive safety-alignment measures have been conducted to armor these models against malicious use (primarily hard prompt attack). However, beneath the seemingly resilient facade of the armor, there might lurk a shadow. (…) these safely aligned LLMs can be easily subverted to generate harmful content. Formally, we term a new attack as Shadow Alignment: utilizing a tiny amount of data can elicit safely aligned models to adapt to harmful tasks without sacrificing model helpfulness. Remarkably, the subverted models retain their capability to respond appropriately to regular inquiries (Yang et al.)

Safety alignments are created to ensure that no harmful, inappropriate, or restricted content is generated. The two main techniques hackers utilise in manipulating or exploiting large language models (LLMs) to shadow-align their built-in content filters and safety mechanisms are known as jailbreaking and prompt injection. Jailbreaking involves exploiting the underlying architecture or loopholes in the model's training data, allowing users to manipulate the model into generating responses that go against its intended guidelines.

Prompt injection is a method employed to manipulate the responses of large language models (LLMs) by incorporating concealed or harmful instructions into apparently harmless input prompts. This approach capitalises on the model's inclination to adhere to provided instructions, thus introducing adversarial directives that have the potential to modify the model's behaviour. As an illustration, a potential intruder could create a prompt that contains concealed commands to retrieve confidential data or execute unauthorised operations. Prompt injection can result in unintended disclosures of private data, the execution of malicious tasks, or the evasion of content moderation systems. These hacking techniques are acquiring particular significance as LLMs become more integrated into different applications and platforms. (Schulhoff et al.)

Effectively injecting the system with a disturbance that doesn't align with the model. Shadow alignment serves as a counteracting force against the normality of shadow prompting, which initially emerged as a means to control the disorderly inclinations of the GenAI phenomenon.

Conclusions

As mentioned in the introduction of this paper, the research conducted by Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal provides valuable insights in support of the autophagic theory. Empirical evidence and theoretical analyses indicate that model collapse is an unavoidable consequence of recursive training, leading to a notable decline in performance over time. This research brings attention to a significant obstacle in the advancement and implementation of generative AI models. As generative models presumably continue to advance and become more integrated into different applications, the occurrence of model collapse demonstrates the need for intelligent data management and the necessity for ongoing access to genuine data sources.

What subversive forms of synthetic media is this autophagic model likely to produce and what impact will it have on culture? Perhaps the answer is noise.

Ting-Chun Liu and Leon-Etienne Kühr in their lecture *Self-cannibalizing AI - Artistic Strategies to expose generative text-to-image models* discuss the feedback loop as an artistic strategy to investigate the latent space of machine learning. Entering their exploration of algorithms for encoding and decoding images within a self-cannibalizing loop using the generative AI model, the outcome yielded a chaotic and indistinct visual representation. Further repetitions in the process revealed that the automated nature of the loop caused the image to lose clarity and definition, resulting in a distorted and abstract final product. Noise.

Their claim is that the presence of feedback loops is an intrinsic characteristic of Stable Diffusion, which is the prevailing model of Text-to-Image AI. When a text is prompted, data is transmitted through a network that generates an image based on probability, spatial arrangement, and quantity. This process resembles an exchange of information between the components, a trial and error process. Drawing on the metaphor of an organism, it is interesting to note that CLIP (Contrastive Language-Image Pre-training) (*CLIP*), one of the primary models created by OpenAI, was initially developed in the medical field to identify tumours in X-ray images.

Perhaps GenAI has indeed caused a metastasis in the systems we utilise to generate and consume online content, or perhaps it is simply a temporary disturbance that will eventually dissipate.

Works cited

- "AUTOPHAGY | English Meaning," *Cambridge Dictionary*. https://dictionary.cambridge.org/dictionary/english/autophagy.
- Bratton, Benjamin. "The Model Is The Message." *Noema Magazine*, July 2022. https://www.noemamag.com/the-model-is-the-message.
- *BUSTA Texts*. https://carolinebusta.github.io/. Accessed 6 May 2024.
- Chang, Natasha C. "Autophagy and Stem Cells: Self-Eating for Self-Renewal." *Frontiers in Cell and Developmental Biology*, vol. 8, Mar. 2020, p. 138. https://doi.org/10.3389/fcell.2020.00138.
- Chayka, Kyle. "Does Monoculture Still Exist on the Internet?" *Vox*, 17 Dec. 2019, https://www.vox.com/the-goods/2019/12/17/21024439/monoculture-algorithm-netflix-spotify
- ---. *Filterworld: How Algorithms Flattened Culture*. First edition, Doubleday, 2024.
- *CLIP: Connecting Text and Images*. https://openai.com/index/clip. Accessed 6 May 2024.
- *Dead Internet Theory: Most of the Internet Is Fake | Agora Road's Macintosh Cafe*. https://f

rum.agoraroad.com/index.php?threads/dead-internet-theory-most-of-the-internet-is-fake.3011/. Accessed 6 May 2024.

- "Detoxifying Cybernetics:From Homeostasis to Autopoiesis and Beyond," *Medialab.* https://medialab.timesmuseum.org/en/lectures/symposium-ii/katherine-hayles. Accessed 31 July 2024.
- Gardels, Nathan. "The Babelian Tower Of AI Alignment." *Noema Magazine,* Apr. 2024. https://www.noemamag.com/the-babelian-tower-of-ai-alignment.
- Jung, C. G. *Mysterium Coniunctionis: An Inquiry into the Separation and* Synthesis of Psychic Opposites in Alchemy. *2d ed, Princeton University Press, 1970.*
- López-Corona, Oliver, and Gustavo Magallanes-Guijón. “It Is Not an Anthropocene; It Is Really the Technocene: Names Matter in Decision Making Under Planetary Crisis.” *Frontiers in Ecology and Evolution,* vol. 8, June 2020, p. 214. https://doi.org/10.3389/fevo.2020.00214
- Metz, Cade, et al. “How Tech Giants Cut Corners to Harvest Data for A.I.” *The New York Times,* 6 Apr. 2024. https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.
- Navaroli, Anika Collier. “Op-Ed: AI’s Most Pressing Ethics Problem.” *Columbia Journalism Review,* https://www.cjr.org/tow_center/op-ed-ais-most-pressing-ethics-problem.php. Accessed 6 May 2024.
- *NEW MODELS 2024®.* https://newmodels.io/. Accessed 6 May 2024.
- Parzych, Katherine R., and Daniel J. Klionsky. “An Overview of Autophagy: Morphology, Mechanism, and Regulation.” *Antioxidants & Redox Signaling,* vol. 20, no. 3, Jan. 2014, pp. 460–73. https://doi.org/10.1089/ars.2013.5371
- Presse, K. W. “KW Digital: Open Secret.” *KW Institute for Contemporary Art,* 8 June 2021, https://www.kw-berlin.de/open-secret/.
- Read, Max. “How Much of the Internet Is Fake?” *Intelligencer,* 26 Dec. 2018, https://nymag.com/intelligencer/2018/12/how-much-of-the-internet-is-fake.html
- Salvaggio, Eryk. “Shining a Light on ‘Shadow Prompting’” *Tech Policy Press,* 19 Oct. 2023, https://techpolicy.press/shining-a-light-on-shadow-prompting.
- Schulhoff, Sander, et al. “Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs Through a Global Prompt Hacking Competition." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* Association for Computational Linguistics, 2023, pp. 4945–77. https://doi.org/10.18653/v1/2023.emnlp-main.302
- *Self-Cannibalizing AI.* Directed by Ting-Chun Liu and Leon-Etienne Kühr, 100AD. media.ccc.de, https://media.ccc.de/v/37c3-12125-self-cannibalizing_ai.
- Shumailov, Ilia, et al. “AI Models Collapse When Trained on Recursively Generated Data.” *Nature,* vol. 631, no. 8022, July 2024, pp. 755–59. https://doi.org/10.1038/s41586-024-07566-y
- “Technosphere Magazine: Home.” *Technosphere Magazine,* https://technosphere-magazine.hkw.de/. Accessed 6 May 2024.
- "These Clues Hint at the True Nature of OpenAI’s Shadowy Q* Project," *Wired* https://www.wired.com/story/fast-forward-clues-hint-openai-shadowy-q-project/. Accessed 6 May 2024.
- Tiffany, Kaitlyn. “Maybe You Missed It, but the Internet ‘Died’ Five Years Ago.” *The Atlantic,* 31 Aug. 2021, https://www.theatlantic.com/technology/archive/2021/08/dead-internet-theory-wrong-but-feels-true/619937/.
- *Why Computer-Made Data Is Being Used to Train AI Models.* https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de. Accessed 6 May 2024.
- Yang, Xianjun, et al. "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models." *arxiv.* https://doi.org/10.48550/ARXIV.2310.02949.

Biography

Luca Cacini is an interdisciplinary artist and researcher in new media. His work investigates the intersections of queer ecology and techno-capitalism. He is part of the Media Arts Cultures Erasmus Mundus Joint Master Degree program at the University for Continuing Education Krems, Aalborg University, University of Lodz, and Lasalle College of the Arts.