On the Caveats of AI Autophagy

Xiaodan Xing[1,2], Fadong Shi[1], Jiahao Huang[1,2], Yinzhe Wu[1,2], Yang Nan[1,2], Sheng Zhang[1,2], Yingying Fang[1,2], Michael Roberts[5,6], Carola-Bibiane Schönlieb[5], Javier Del Ser[7,8], Guang Yang[1,2,3,4]

[1*]Bioengineering Department and Imperial-X, Imperial College London, London, UK.
[2]National Heart and Lung Institute, Imperial College London, London, UK.
[3]Cardiovascular Research Centre, Royal Brompton Hospital, London, UK.
[4]School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK.
[5]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK.
[6]Department of Medicine, University of Cambridge, Cambridge, UK.
[7]Department of Communications Engineering, University of the Basque Country UPV/EHU, Bilbao, Spain.
[8]TECNALIA, Basque Research and Technology Alliance (BRTA), Bilbao, Spain.

Abstract

Generative Artificial Intelligence (AI) technologies and large models are producing realistic outputs across various domains, such as images, text, speech, and music. Creating these advanced generative models requires significant resources, particularly large and high-quality datasets. To minimise training expenses, many algorithm developers use data created by the models themselves as a cost-effective training solution. However, not all synthetic data effectively improve model performance, necessitating a strategic balance in the use of real versus synthetic data to optimise outcomes. Currently, the previously well-controlled integration of real and synthetic data is becoming uncontrollable. The widespread and unregulated dissemination of synthetic data online leads to the contamination of datasets

traditionally compiled through web scraping, now mixed with unlabeled synthetic data. This trend, known as the AI autophagy phenomenon, suggests a future where generative AI systems may increasingly consume their own outputs without discernment, raising concerns about model performance, reliability, and ethical implications. What will happen if generative AI continuously consumes itself without discernment? What measures can we take to mitigate the potential adverse effects? To address these research questions, this study examines the existing literature, delving into the consequences of AI autophagy, analyzing the associated risks, and exploring strategies to mitigate its impact. Our aim is to provide a comprehensive perspective on this phenomenon advocating for a balanced approach that promotes the sustainable development of generative AI technologies in the era of large models.

Keywords: Generative AI, AI Autophagy, Large Language Models, Watermarking.

1 Introduction

Deep learning-based generative models have significantly transformed AI and machine learning. These models can generate highly realistic data, with applications in creating images, text, speech, and music. Innovations like DALL-E 2 [1], Imagen [2], Suno [3], and Sora [4], alongside advanced language models like ChatGPT [5], demonstrate their vast potential in diverse fields such as art, entertainment, medical diagnostics, and automated content generation. These technologies enable seamless integration of visual, textual, and auditory data, showcasing their broad applicability.

Training these deep generative models is not a cheap task. They require extensive, high-quality datasets, leading many organisesations to scrape web data. For instance, CommonCrawl [6], a petabyte-scale open-source web crawling database, is widely used to train generative models such as StableDiffusion and GPT-3. This dataset is regularly expanded with 3–5 billion new pages each month, accounting so far for a total of 250 billion website pages.

As web scraping increases, so does the risk of incorporating synthetic data generated by other models. This means models could be trained on data produced by previous generative models. Currently, most generative models are trained on human-generated content. However, with more synthetic data online, these models may start training on artificial rather than original data. Additionally, the rapid proliferation of data-intensive models is depleting high-quality, human-generated data. Projections suggest that by 2026, the pool of suitable language data for training large language models (LLMs) may be depleted [7]. As AI-generated outputs grow, including AI-generated content in training datasets seems inevitable.

Initially, training generative AI on synthetic content might seem beneficial, as integrating the right amount can yield impressive outcomes [8]. However, as models evolve over generations, researchers have flagged potential negative impacts [9–11]. This phenomenon, where generative models are trained on AI-generated content, is termed "autophagy" [10], similar to the biological process where a cell consumes itself.

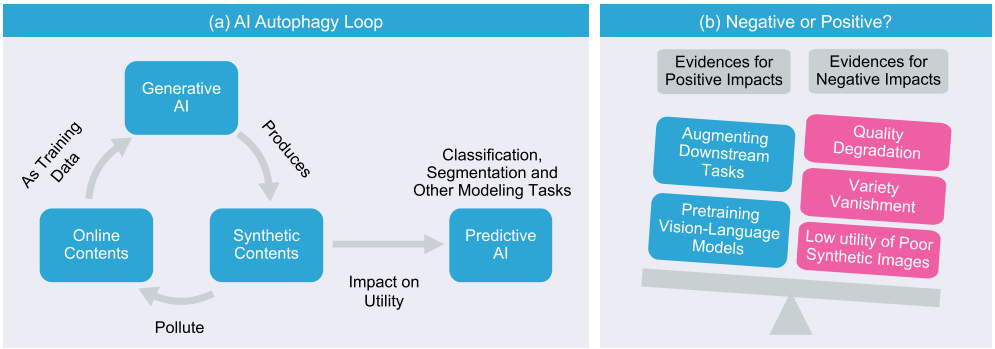

Fig. 1 A diagram showing (a) the scope of this work and (b) the major research questions. This study centers on the AI autophagy phenomenon, with a focus on analyzing its effects and discussing solutions to mitigate potential negative impacts.

Performance degradation due to synthetic data is a concern beyond generative models. In computer vision, tasks like classification and segmentation can suffer when using synthetic data. Literature shows that fully synthetic data can degrade performance compared to models trained on real data [11, 12]. As models consume synthetic online content, potential consequences in practical settings, especially high-risk scenarios, are significant [13].

Despite the wide recognition of AI autophagy in recent times, a significant gap remains in comprehensive studies that explore both theoretical and empirical analyses of this phenomenon. Albeit widely discussed in the literature, findings often conflict, leading to divided opinions on the impact of synthetic data on AI models. Some studies suggest that synthetic data pollution might not pose a major concern under controlled conditions: the models' performance will stop deteriorate if carefully managed [14, 15]. In contrast, others present experimental evidence proving the severe and unavoidable consequences of AI autophagy, emphasizing the potential performance deterioration of the model resulting therefrom [10, 16–18].

Moreover, there is a clear lack of practical analyses that integrate both technical and regulatory viewpoints. Most existing research tends to focus on isolated technical solutions [19, 20], overlooking the need for a comprehensive approach that examines how these strategies can be effectively implemented in real-world scenarios, particularly within existing worldwide regulatory frameworks.

As shown in Figure 1, this work aims to fill the research gap by providing an integrated study that critically examines empirical, theoretical, technical, and regulatory evidence, offering definitive insights and practical recommendations for managing synthetic data in AI development. The objective of this paper is to address three research questions:

- **RQ1:** What are the consequences of out-of-control autophagy with synthetic contents?
- **RQ2:** What technical strategies can be employed to mitigate the negative consequences of AI autophagy?

- **RQ3:** Which regulatory strategies can be employed to address these negative consequences?

2 RQ1: What Happens When AI Eats Itself?

The incorporation of synthetic data into training sets is inevitable [7] and may reshape our understanding of AI's capabilities and limitations. In this section, we discuss the potential consequences of this trend, specifically examining whether models improve or deteriorate when synthetic data is recursively added to the training dataset.

2.1 Fully Synthetic Loop: Worst Case and Theoretical Model

In the worst-case scenario, the generative model is trained predominantly on its own outputs. This situation is known to distort the data distribution, resulting in a marked decline in model performance [9, 10, 21]. While a fully synthetic loop is not typical in real-world applications, examining this scenario sheds light on the theoretical understanding of **model collapse**—a phenomenon where AI autophagy leads to severe performance degradation.

A simple and intuitive way to understand the consequences of a fully synthetic loop is through a Gaussian distribution. The generative process involves sampling data points from a reference Gaussian distribution and using these samples to approximate the original distribution's characteristics. In a fully synthetic loop, each generative model only has access to data generated by previous models.

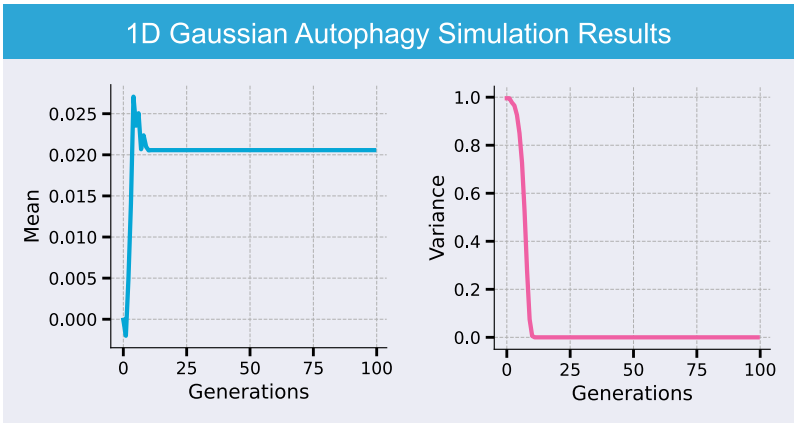

Fig. 2 The trend of mean and variance estimations from the autophagy loop. Here we chose the starting value $\mu = 0$ and $\sigma = 1$. For each round, 10^4 data points were generated to perform the estimation.

Through this iterative process, as shown theoretically in [9, 10] and in the experimental results in Figure 2, we observe how the mean and variance evolve across multiple rounds of autophagy: (1) the mean values undergo a random walk, signifying potential degradation in data quality, and (2) the variance decreases over time, leading to a loss of diversity in the generated outputs.

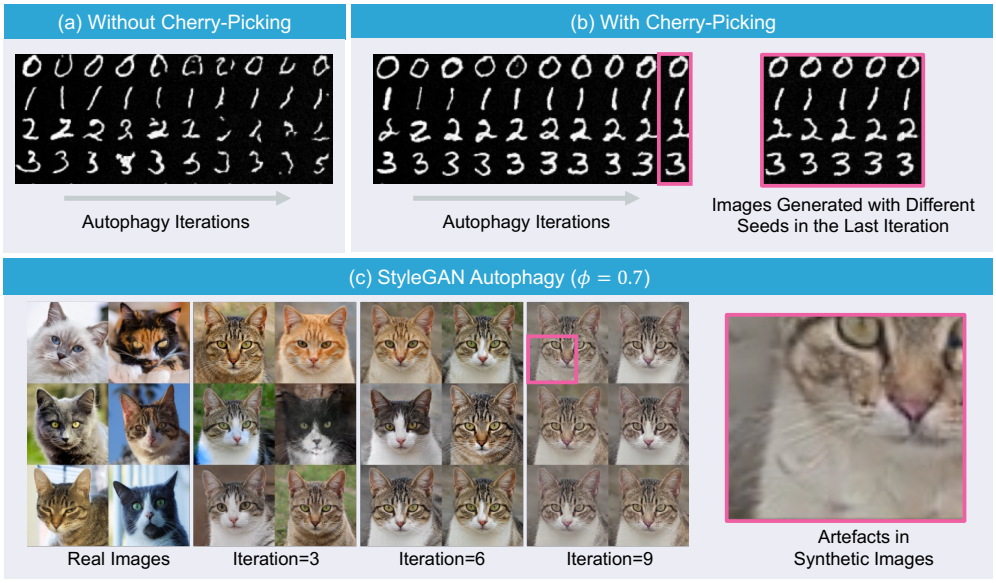

Fig. 3 Synthetic images generated from an autophagy loop example using a DDPM model on the MNIST dataset [24] illustrate quality degradation (a). In (b), applying a selective process that retains only high-quality images for subsequent autophagy loops reduces diversity. (c) presents results from a more complex task using the StyleGAN model [25] on the AFHQ cat dataset [26], with a truncation parameter of 0.7 to balance diversity and quality. Despite these adjustments, quality degradation and diversity loss remain evident during autophagy

Based on this simple model, Bertrand et al. [14] extended the analysis to multivariate Gaussian distributions, finding that larger sample sizes in each iteration slow the rate of collapse.

To investigate how errors propagate through the model's predictions over time, Dohmatob et al. [22] employed a linear regression model, which offers a clear framework for understanding how errors introduced in each iteration of training can compound over time, leading to significant deviations in the model's predictions. In this setup, they discovered that the test error grows linearly with the number of iterations in fully synthetic loop.

Empirical findings from both images shown in Figure 3(a) and language data support these theoretical conclusions. For example, Alemohammad et al. [10] observed that degradation in the StyleGAN2 model appears as cross-hatched artifacts, while Martínez et al. [16, 17] reported fuzziness and blurring in the DDIM model during autophagy. Bohacek et al. [21] identified repetitive pixels and patterns in the Stable Diffusion Model. Similarly, in language models, Shumailov et al. [9] found that the autophagy loop reduces output diversity, as shown in Figure 4. Additionally, Guo et al. [23] analysed diversity metrics across iterations, revealing a consistent decline in output diversity, particularly for tasks requiring high levels of creativity.

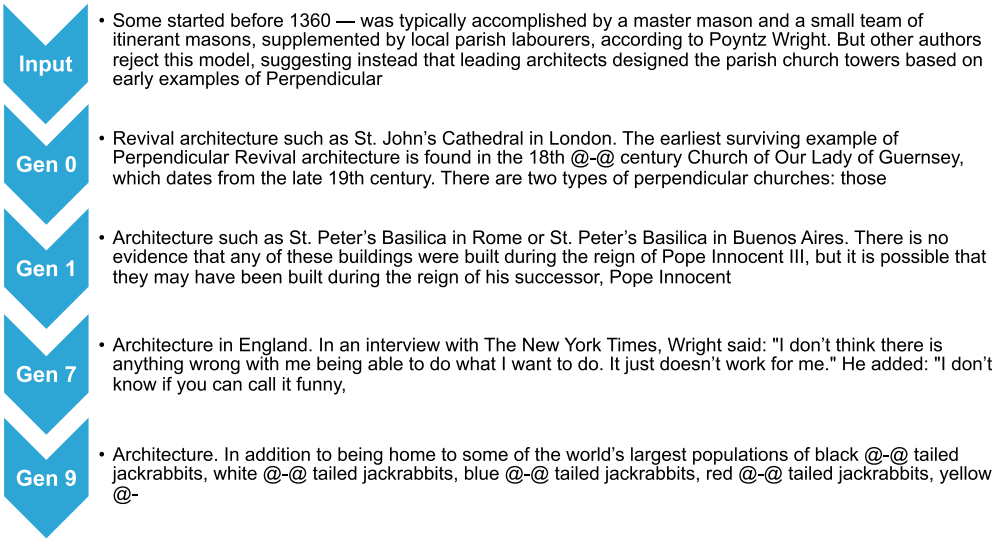

Fig. 4 Synthetic texts from an autophagy loop from the study in [9], demonstrating a reduction in text variety. In later generations, the texts increasingly exhibit duplication.

2.2 Fixed-Real Data Loop: Stability and Limitations

Introducing a fixed real dataset can create a **fixed-real data loop**, where AI-generated images are incrementally combined with real data. In each iteration, a fixed real dataset is incorporated.

Theoretically, Bertrand et al. [14] demonstrated the stability of iterative retraining when a fixed real dataset is preserved, i.e., if a fixed real data is observed, the quality degradation will become stable and not change. The study proposed a theoretical framework for iterative retraining of generative models, building upon the standard maximum likelihood objective. Fu et al. [27] also describes a theoretical guarantee on the stability and error bounds when training a model with a mixture of real and synthetic data. However, the stability in [14] and [27] relies on unfeasible settings: (1) the initial generative model is sufficiently well-trained, and (2) each retraining iteration retains a high proportion of the original real data.

Gerstgrasser et al. [15] observed similar results using traceable linear models, where the test error stabilises at a finite upper bound as the number of iterations increases, rather than continuing to grow. Nevertheless, this is still problematic: if the model has already collapsed into an unacceptable level of performance, even if the error ceases to worsen, the outcome remains unsatisfactory.

Empirically, keeping real data in the loop also slows the rate of model collapse [10, 15–18], it does not solve the problem entirely.

When real data is retained and synthetic data is scaled up, normal scaling laws break down. Typically, scaling laws predict performance improvements with more training data. However, mixing synthetic data with real data changes the data distribution, breaking the expected scaling law [28].

2.3 Fresh Data Loop: A Temporary Solution or a Real Fix?

A fresh data loop introduces new real data at each iteration. Studies [10, 18] have shown that incorporating fresh real data can temporarily mitigate quality degradation. [21] also demonstrated that adding fresh real data can partially heal the data loop.

However, while the fresh data loop can **slow down** or **delay** degradation, it does not entirely resolve the underlying issue. The proportion of real data versus synthetic data becomes a critical factor: as the dataset grows, if the amount of synthetic data significantly outweighs the fresh real data being introduced [29], the model may still struggle to generalise effectively. This imbalance can still lead to performance stagnation or eventual collapse, albeit at a slower rate.

Moreover, the practical feasibility of consistently obtaining fresh real data is a challenge. In many real-world applications, collecting new high-quality data at every iteration is either expensive or impractical. Thus, while the fresh data loop offers a temporary buffer against the degenerative effects of synthetic data, it is not a complete solution.

2.4 The Diminished Role of Synthetic Data in Data Augmentation

Before concerns arose about the AI autophagy loop, synthetic data was widely praised for its utility in augmenting training datasets, particularly for downstream tasks such as classification and image generation [29, 30]. Synthetic data was recognised for improving sample variety and providing valuable pre-trained parameters. However, recent studies have raised doubts about its effectiveness.

Ravuri and Vinyals [31] found that mixing generated samples with real data often degrades classifier accuracy. Although using low truncation values in GANs can improve accuracy when only a small amount of generated data is used, performance drops significantly when the proportion of synthetic data increases beyond a certain threshold, falling below the performance of models trained exclusively on real data.

Hataya et al. [12] empirically addressed this issue by simulating dataset contamination. They generated ImageNet-scale and COCO-scale datasets using state-of-the-art generative models and evaluated the impact of "contaminated" datasets on various tasks, including image classification and generation. Their findings showed that generated images negatively affect downstream performance, with the degree of degradation depending on the ratio of synthetic images to real images and the specific downstream task. Further analysis revealed that generated images also reduce robustness to out-of-distribution data, indicating that the use of synthetic data for augmentation poses a threat to the safety of AI models in open-world environments.

Chen et al. [32] further highlighted the risks by focusing on dataset bias, showing that even well-generated images tend to amplify existing biases in the data.

It is noteworthy noting that these findings have not accounted for autophagy loops, where synthetic data is reintroduced into the training cycle. Still, they offer a pessimistic view of the future impact of synthetic data on AI-based systems.

3 RQ2: What Technical Strategies Can Be Employed to Mitigate the Negative Consequences of AI Autophagy?

3.1 The Limitations of Cherry-Picking in Mitigating AI Autophagy

A straightforward solution to mitigate model collapse is "cherry-picking"—removing low-quality or incorrect synthetic contents in generative models to ensure only high-quality, realistic ones are used in subsequent iterations of training. Techniques like the truncation trick in GANs [33] and guidance parameters in diffusion models also help generate higher-quality images. Similarly, in language models, hallucinated or linguistically unacceptable content can be detected and filtered out [23] to maintain quality.

However, these methods often worsen model collapse by reducing diversity, as noted in [10], as as shown in Figure 3 (b) and (c). A similar issue arises in language generation models. For example, Guo et al. [23] used a linguistic acceptability filter to remove noisy samples during each iteration, discarding the 20% of synthetic data with the lowest acceptability scores before retraining the model. While this approach improves immediate data quality, it exacerbates model collapse by narrowing the model's generative variety. In contrast, Shumailov et al. [9] introduced a penalty to reduce repetition, which helped mitigate model collapse. However, as a result of this penalty, models began generating lower-quality sentences in an effort to avoid repetition, ultimately reducing the overall performance of the model.

Researchers are also exploring feedback mechanisms to filter suboptimal samples [19, 20, 34]. Feng et al. [19] showed that feedback-augmented synthetic data, such as pruning incorrect predictions or selecting the best guesses, can prevent model collapse, aligning with techniques like Reinforcement learning from human feedback (RLHF). Similarly, Ferbach et al. [20] demonstrated that curating data based on a reward model maximises expected rewards, suggesting that increasing curated synthetic data optimises preferences for future models. However, these methods fail to fully resolve the inherent trade-off between quality and variety in the generative process, overlooking the challenge of designing a universal reward system for synthetic data selection in general-purpose AI.

In addition to the previously discussed cherry-picking method, retaining real data in the training cycle has been both theoretically and empirically proven to mitigate the effects of model collapse [10, 18]. Additionally, [35] proposed using synthetic datasets to train an auxiliary model that adjusts the diffusion models in a negative direction, helping to prevent quality degradation during autophagy iterations. However, it is important to recognize that the success of these methods fundamentally relies on the accurate identification of synthetic data. Thus:

> We must always be able to discern which parts of the training data are synthesised and which are human created.

3.2 Advantages and Caveats of Synthetic Content Watermarking

Watermarking is one way to help users identify the synthetic contents [36–43].

Watermarking synthetic images is a promising approach, already finding real-world applications in AI-driven image generation. In 2023, Google DeepMind and Google Cloud launched SynthID, a tool for watermarking and identifying AI-generated images. SynthID embeds an invisible digital watermark in an image's pixels, detectable for verification. It was released to select Vertex AI customers using Imagen, a text-to-image model. StableDiffusion also includes an invisible watermarking feature in its reference sampling script, using the Python package invisible-watermark with DCT and DWT algorithms.

Watermarking in language models has limited practical applications. In real-world use, users typically iteract with black-box LLMs, making it impractical to embed watermarks within the model's generation process. Post-generation watermarking, which alters output texts, is less effective, often degrading text quality and remaining vulnerable to paraphrasing attacks [44, 45].

Robustness to attacks is a critical concern in watermarking methods. It is important to emphasise that our focus is not on extreme scenarios involving sophisticated, malicious adversarial attacks, but rather on ensuring that watermarks remain resilient against unintended attacks, perturbations and modifications that occur naturally, such as JPEG compression, cropping, and rephrasing of text. Robustness against these everyday alterations is a key quality metric in many technical studies [41, 43, 44, 46, 47]. With proper stakeholder cooperation, current watermarking techniques are generally effective at resisting these simple distortions. However, this raises another practical question [43]: if a human extensively rephrases AI-generated text, can it still be considered "AI-generated"? This issue is largely overlooked in the current literature, and there are no clear boundaries by which a reworded text remains AI-generated, nor the extent by which a reworded text used to update a generative model can contribute to the degradation and mode collapse described previously.

Furthermore, effective watermark detection often depends on access to technical details that companies are typically unwilling to share, creating additional barriers to the effective implementation and reliability of watermarking systems.

3.3 Advantages and Caveats of Synthetic Content Detection

In addition to watermarks, researchers are exploring passive detection (or forensics detection) algorithms [48] based on the inherent differences between synthetic and human-generated content [49–56]. These algorithms do not require direct interaction with the AI system that created the content.

Unlike imaging data, non-watermark-based detection techniques are prevalent for AI-generated language on commercial platforms. These techniques, such as GPTZero [53], Originality.ai [57], and Turnitin's AI detection tools [58], are non-intrusive and do not compromise text quality.

However, a significant challenge for non-watermark based LLM detectors lies in their generalisability, or their capacity to recognise AI-generated texts from new models. As AI technology progresses, the distinction between human and AI-generated content is becoming increasingly blurred. For instance, in evaluations using a "challenge set" composed of high-quality synthetic and human-generated English texts, OpenAI's text classifier correctly identified only 26% of the generated content [52].

Concerns about the accuracy of synthetic text detectors go beyond precision, highlighting issues of explainability and fairness. Research [59] shows that leading AI detectors often misclassify writings by non-native speakers as AI-generated. Some universities, like Vanderbilt [60], have stopped using these detectors because companies refuse to disclose their identification methods.

4 RQ3: Which Regulatory Strategies Can Be Employed to Address These Negative Consequences?

Section 3 explores technical solutions to tackle the problem of AI autophagy, emphasizing the importance of marking and detecting synthetic content. This approach is identified as the most practical and straightforward method to address the issue. Integrating technical and regulatory efforts is crucial to ensure the success of these technological approaches.

4.1 Problematic Data Acquisition and Dissemination Strategies

Current policies of online generative platform companies have failed to significantly address the AI autophagy loop. This failure is evident in their unregulated absorption of online data without implementing strategies to filter synthetic content [61–64]. While these companies have adopted responsible AI measures to limit the inclusion and dissemination of harmful or unethical content [62, 64], they have not specifically addressed the issue of synthetic content. Current strategies primarily focus on reducing explicit, violent, or biased content during distribution, rather than marking and filtering synthetic synthetic content from training processes, leaving a significant gap in combating the AI autophagy loop.

Although solid evidence directly linking these problematic policies to negative impacts is limited, some indications of their effects include performance degradation in models such as GPT-3.5 and GPT-4 [65]. These models exhibited increased formatting errors in code generation from March to June 2023 [65]. While the exact causes are not fully clear, it is speculated that the pollution of training datasets with AI-generated data could be a contributing factor [66].

4.2 Relevant Regulations for the Dissemination Process

Combating the AI autophagy loop requires government regulations to ensure that synthetic content is identifiable and detectable, facilitating its filtering from training datasets. This transparency is essential not only for addressing the AI autophagy

loop but also for broader concerns such as preventing the spread of misinformation, protecting intellectual property, and maintaining trust in digital media.

China: China has issued the *Interim Measures for the Management of Generative AI Services*, requiring providers to clearly mark generated content, such as images and videos, to prevent public confusion (Article 12) [67]. Similarly, the *Artificial Intelligence Act of 2021* mandates that users of AI systems generating realistic content, such as deepfakes, must label such content as artificially generated or manipulated.

European Union (EU): The *EU AI Act (2024)* categorises AI systems based on their risk levels [68, 69]. AI systems must comply with information and transparency requirements. Users must be informed when they are interacting with AI, such as chatbots. Deployers of AI systems that generate or manipulate images, audio, or video content, including deepfakes, are required to disclose that the content has been artificially generated or manipulated, except in limited cases such as for the prevention of criminal offenses. Providers of AI systems generating large quantities of synthetic content must implement reliable, interoperable, and robust techniques, such as watermarks, to mark and identify AI-generated or manipulated content.

United States (US): The US has introduced a presidential regulation focused on addressing AI risks. The regulation calls for the development of effective labeling and content provenance mechanisms, enabling Americans to determine when content has been generated or manipulated by AI [70]. These actions are intended to address AI's risks while preserving its potential benefits.

While it is still early to fully assess the impact of these regulations, they have already begun to influence the global conversation on AI governance and accountability [71]. Regulations are crucial for controlling the misuse of synthetic data, but relying solely on the integrity of synthetic data disseminators is impractical [72]. Without effective detection methods for identifying synthetic data, ensuring compliance with these laws remains a significant challenge.

5 Conclusions and Outlook

In this paper, we revisit the concept of AI autophagy, where generative models iteratively consume their own synthetic outputs. Our perspective covered theoretical insights, experimental findings, and potential technical and regulatory solutions to address this phenomenon. Below, we summarise the findings, structured around the original questions posed in the introduction:

- **RQ1: What are the consequences of uncontrolled autophagy with synthetic contents?**

When AI models consume their own outputs—a phenomenon known as AI autophagy—the data distribution becomes increasingly distorted [9, 28], leading to significant performance degradation over time. Iterative training on synthetic data causes a gradual decline in both quality and diversity of generated outputs, as demonstrated by both theoretical models, including Gaussian [9, 10, 14], probabilistic [14, 27], and linear [15, 22] frameworks, and empirical findings [9, 10, 16, 17, 21, 23, 28].

Incorporating fixed-real data [14, 15] or introducing fresh data [10], as shown in Figure 5, into training cycles offers only temporary relief. While these strategies can slow the rate of collapse, they do not address the fundamental issue, as maintaining a balance between synthetic and real data is challenging, particularly in real-world scenarios where high-quality real data is often scarce.

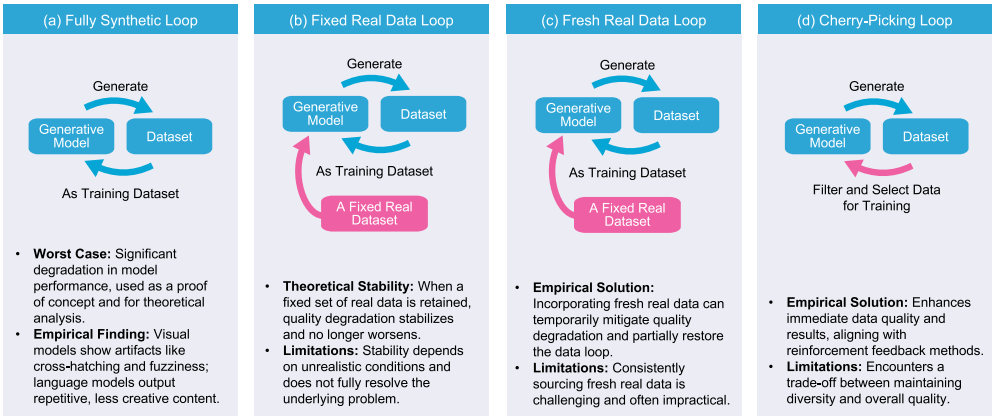

Fig. 5 Different loops in AI autophagy.

- **RQ2: What technical strategies can be employed to mitigate these negative consequences?**

While strategies like incorporating real data into the training process and filtering out low-quality synthetic data can mitigate these issues, these approaches are constrained by practical limitations. Incorporating real data requires consistent sourcing, which is often impractical, and filtering strategies can inadvertently reduce data diversity, exacerbating long-term model collapse. Moreover, the effectiveness of these mitigation methods fundamentally depends on accurately distinguishing between real and synthetic data—a task that grows increasingly complex as synthetic data becomes more sophisticated.

Common techniques for distinguishing synthetic content include watermarking and detection methods. While watermarking is cost-effective, it can impact content quality and is not universally applicable. Detection methods avoid these drawbacks but struggle with generalisesation and transparency. Implementing widespread protection requires collaboration from regulatory authorities and commercial entities, but universal strategies risk uniform attacks. Additionally, concerns about the accuracy, explainability [60] and fairness [59] of synthetic content detectors persist.

- **RQ3: Which regulatory strategies can be employed to mitigate these negative consequences?**

Current policies implemented by generative AI companies are insufficient, highlighting the need for stronger collaboration between regulatory bodies and corporations.

Regulations must work in tandem with industry efforts to develop effective techniques for identifying AI-generated content, which can then be filtered out from training datasets. This approach would not only address the AI autophagy loop but also mitigate broader concerns, such as the spread of misinformation, intellectual property violations, and the erosion of trust in digital content.

Current regulatory strategies proposed include mandating the marking of generated content to prevent public confusion, requiring real-name verification for content creators to enhance traceability, and enforcing clear labeling of AI-generated content, particularly deep fakes. However, the effectiveness of these strategies relies on the development of reliable detection methods for AI-generated content, demonstrating a need for a collaborative effort among technology companies, regulatory agencies, and civil society to forge effective governance strategies, safeguarding the web's content integrity.

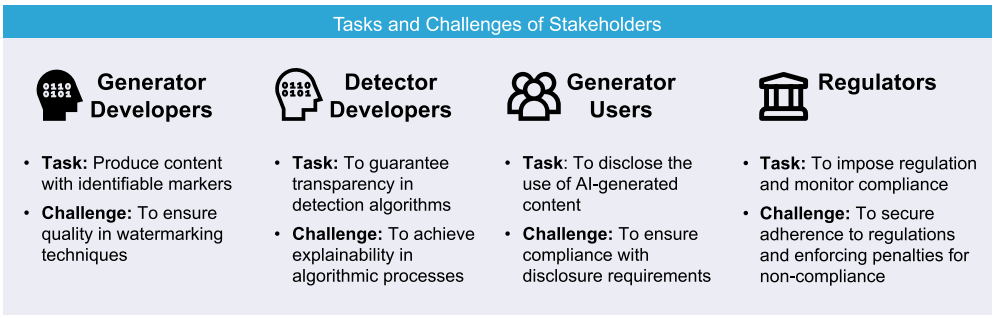

Fig. 6 Overview of tasks and challenges faced by different stakeholders in the context of preventing data pollution by generative models.

A key takeaway from our paper is that no single technical solution stands out as a definitive answer to addressing the AI autophagy issue. While maintaining a certain amount of real data and filtering out high quality synthetic data for training serves as a temporary fix, its effectiveness still relies on our ability to distinguish between synthetic and real data accurately.

As a summarisation of this work, we present in Figure 6 an overview of the roles and corresponding challenges faced by various stakeholders in the context of AI-generated content. **Generator developers** are individuals or organizations, including labs or large companies, that create and refine deep generative models. Their key task is to embed identifiable markers, like watermarks, in AI-generated content to distinguish it from human-made material, without degrading quality—requiring sophisticated techniques. **Detector developers**, which may include governments and authorities, focus on building tools to detect AI-generated content. Their role is vital for verifying authenticity, and they must ensure detection systems are transparent and trustworthy. Their challenge is balancing explainability with maintaining the accuracy and robustness of their systems. **Generator users**—such as content creators, businesses, and researchers—use AI-generated content and are responsible for disclosing its use.

They play a crucial role in ensuring transparency to prevent misinformation, but complying with disclosure requirements can be difficult, especially if regulations are not strictly enforced. **Regulators**—government agencies, international bodies, or industry groups—create and enforce rules for AI-generated content. They are responsible for mandating watermarking, detection, and disclosure, and for protecting the public from risks like misinformation. Their challenge is keeping regulations enforceable in a rapidly changing AI landscape.

With the responsibilities and challenges clearly identified among these stakeholders, we can focus on building a comprehensive framework that enables collaboration across sectors to prevent the risks associated with AI-generated content. By working together, generator developers, detector developers, generator users, and regulators can address the shortcomings of current technical solutions, such as watermarking and detection tools, and ensure that these technologies are robust and scalable in real-world applications. This collaborative framework will also strengthen the enforcement of regulations, helping to mitigate issues like misinformation, data pollution, and non-compliance, which are critical in a rapidly evolving AI landscape.

6 Ethical and Societal Considerations

The aim of this study is to assess the potential adverse effects of autophagy loops in generative models, particularly the impact of synthetic data proliferation across the web. These findings highlight the risks associated with unchecked AI-generated content, such as degradation in model quality, misinformation, and potential harm to data integrity in various fields.

It is important to emphasize that this concern does not imply endorsement of practices such as large-scale web scraping for training generative AI models. The ethical implications of data sourcing, especially without consent, pose serious challenges to privacy and data ownership, and should be carefully regulated.

The intended use of the findings presented in this paper is to encourage collaborative efforts among generator developers, detector developers, users, and regulators to create frameworks that promote transparency, accountability, and responsible AI use. It is our hope that these solutions can help mitigate risks such as data pollution and ensure AI-generated content is used ethically and for the benefit of society.

This work also acknowledges that while technological solutions, such as watermarking and detection systems, are promising, they are not without limitations. Ensuring compliance with ethical standards in real-world applications requires robust enforcement mechanisms, cooperation among stakeholders, and ongoing dialogue with policymakers and civil society.

7 Acknowledgements

This study was supported in part by the ERC IMI (101005122), the H2020 (952172), the MRC (MCPC21013), the Royal Society (IEC\NSFC\211235), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd, the UKRI Future Leaders Fellowship (MRV0237991), and Wellcome Leap's Dynamic Resilience programme, jointly funded with Temasek Trust. This work

was made possible through the support of the ARC project team at The Alan Turing Institute, under the project designation ARC-001. J. Del Ser acknowledges funding support from the Basque Government through EMAITEK/ELKARTEK grants, as well as the consolidated research group MATHMODE (IT1456-22).

Correspondence should be sent to G. Yang (g.yang@imperial.ac.uk).

8 Competing interest statement

All authors declare no competing interests.

References

[1] OpenAI: DALL-E 2. https://www.openai.com/dall-e-2. Accessed: 2024-10-20

[2] Research, G.: Imagen. https://imagen.research.google. Accessed: 2024-10-20

[3] AI, S.: Suno AI. https://suno.ai. Accessed: 2024-10-20

[4] AI, S.: Sora AI. https://sora.ai. Accessed: 2024-10-20

[5] OpenAI: ChatGPT. https://www.openai.com/chatgpt. Accessed: 2024-10-20

[6] Common Crawl: Overview. Accessed: March 17, 2024 (2008). https://commoncrawl.org/overview

[7] Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M., Ho, A.: Will we run out of data? an analysis of the limits of scaling datasets in machine learning. arXiv Preprint arXiv:2211.04325 (2022)

[8] Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**(1), 1–48 (2019)

[9] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y.: AI models collapse when trained on recursively generated data. Nature **631**(8022), 755–759 (2024). Highlighted as a highly relevant theoretical analysis and empirical finding on AI autophagy by fine-tuning language models, introducing the term "model collapse".

[10] Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A.I., Babaei, H., LeJeune, D., Siahkoohi, A., Baraniuk, R.G.: Self-consuming generative models go mad. arXiv Preprint arXiv:2307.01850 (2023). Highlighted as a highly relevant theoretical analysis and empirical finding on AI autophagy, introducing the term "autophagy" using image synthesis models.

[11] Yamaguchi, S., Fukuda, T.: On the limitation of diffusion models for synthesizing training datasets. arXiv Preprint arXiv:2311.13090 (2023)

[12] Hataya, R., Bao, H., Arai, H.: Will large-scale generative models corrupt future datasets? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20555–20565 (2023)

[13] Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., Prado, M.L., Herrera-Viedma, E., Herrera, F.: Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation. Information Fusion **99**, 101896 (2023)

[14] Bertrand, Q., Bose, A.J., Duplessis, A., Jiralerspong, M., Gidel, G.: On the stability of iterative retraining of generative models on their own data. arXiv preprint arXiv:2310.00429 (2023). Highlighted as a theoretical analysis and potential solution, proposing the stability of training loops that incorporate fixed real data.

[15] Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., et al.: Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. arXiv preprint arXiv:2404.01413 (2024). Highlighted as a potential solution. Demonstrates that, using a linear regression model, if a fixed set of real data is maintained, the test error stabilizes with a finite upper bound, independent of the number of iterations.

[16] Martínez, G., Watson, L., Reviriego, P., Hernández, J.A., Juarez, M., Sarkar, R.: Combining generative artificial intelligence (ai) and the internet: Heading towards evolution or degradation? arXiv Preprint arXiv:2303.01255 (2023). Highlighted as a highly relevant empirical finding on AI autophagy, demonstrated through image synthesis models.

[17] Martínez, G., Watson, L., Reviriego, P., Hernández, J.A., Juarez, M., Sarkar, R.: Towards understanding the interplay of generative artificial intelligence and the internet. arXiv Preprint arXiv:2306.06130 (2023). Highlighted as a highly relevant empirical finding on AI autophagy, demonstrated through image synthesis models.

[18] Briesch, M., Sobania, D., Rothlauf, F.: Large language models suffer from their own output: An analysis of the self-consuming training loop. arXiv preprint arXiv:2311.16822 (2023). Highlighted as a highly relevant theoretical analysis.

[19] Feng, Y., Dohmatob, E., Yang, P., Charton, F., Kempe, J.: Beyond model collapse: Scaling up with synthesized data requires reinforcement. arXiv preprint arXiv:2406.07515 (2024). Highlighted as a potential solution, proposing pruning incorrect predictions and selecting optimal guesses from multiple outputs, to counteract model collapse when scaling with synthesized data.

[20] Ferbach, D., Bertrand, Q., Bose, A.J., Gidel, G.: Self-consuming generative models with curated data provably optimize human preferences. arXiv preprint

arXiv:2407.09499 (2024). Highlighted as a potential solution to AI autophagy. This paper demonstrates that when data is curated according to a reward model, the expected reward of the iterative retraining process is maximized.

[21] Bohacek, M., Farid, H.: Nepotistically trained generative-ai models collapse. arXiv preprint arXiv:2311.12202 (2023). Highlighted as a highly relevant empirical finding on AI autophagy, demonstrated through image synthesis models.

[22] Dohmatob, E., Feng, Y., Kempe, J.: Model collapse demystified: The case of regression. arXiv preprint arXiv:2402.07712 (2024). Highlighted as a highly relevant theoretical analysis of AI autophagy using a linear regression model, analyzing the accumulation of error in fully synthetic loops.

[23] Guo, Y., Shang, G., Vazirgiannis, M., Clavel, C.: The curious decline of linguistic diversity: Training language models on synthetic text. arXiv preprint arXiv:2311.09807 (2023). Highlighted as a highly relevant empirical finding on AI autophagy, demonstrating the decline of diversity across multiple aspects.

[24] Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine **29**(6), 141–142 (2012)

[25] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)

[26] Choi, Y., Uh, Y., Yoo, J., Ha, J.-W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8188–8197 (2020)

[27] Fu, S., Zhang, S., Wang, Y., Tian, X., Tao, D.: Towards theoretical understandings of self-consuming generative models. arXiv preprint arXiv:2402.11778 (2024). Highlighted as a theoretical support for fixed real dataset loops, providing a theoretical guarantee on stability and error bounds when training models with a combination of real and synthetic data.

[28] Dohmatob, E., Feng, Y., Yang, P., Charton, F., Kempe, J.: A tale of tails: Model collapse as a change of scaling laws. In: Proceedings of the 41st International Conference on Machine Learning (ICML) (2024). Highlighted because this study reveals how scaling with synthetic data can disrupt the scaling laws of large models, leading to model collapse.

[29] Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023)

[30] Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models. arXiv preprint arXiv:2302.07944 (2023)

[31] Ravuri, S., Vinyals, O.: Classification accuracy score for conditional generative models. Advances in neural information processing systems **32** (2019)

[32] Chen, T., Hirota, Y., Otani, M., Garcia, N., Nakashima, Y.: Would deep generative models amplify bias in future models? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10833–10843 (2024)

[33] Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. arXiv Preprint arXiv:1809.11096 (2018)

[34] Gillman, N., Freeman, M., Aggarwal, D., Hsu, C.-H., Luo, C., Tian, Y., Sun, C.: Self-correcting self-consuming loops for generative model training. arXiv preprint arXiv:2402.07087 (2024). Highlighted as a potential solution, proposing using a predefined criteria to filter out incorrect synthetic contents.

[35] Alemohammad, S., Humayun, A.I., Agarwal, S., Collomosse, J., Baraniuk, R.: Self-improving diffusion models with synthetic data. arXiv preprint arXiv:2408.16333 (2024). Highlighted as a solution for AI autophagy by using synthetic datasets to optimise the diffusion model's score function through a fine-tuning step.

[36] Wen, Y., Kirchenbauer, J., Geiping, J., Goldstein, T.: Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. arXiv preprint arXiv:2305.20030 (2023)

[37] Fernandez, P., Couairon, G., Jégou, H., Douze, M., Furon, T.: The stable signature: Rooting watermarks in latent diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision 2023, pp. 22466–22477 (2023)

[38] Bui, T., Agarwal, S., Collomosse, J.: Trustmark: Universal watermarking for arbitrary resolution images. arXiv preprint arXiv:2311.18297 (2023)

[39] Tancik, M., Mildenhall, B., Ng, R.: Stegastamp: Invisible hyperlinks in physical photographs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2117–2126 (2020)

[40] Abdelnabi, S., Fritz, M.: Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In: 2021 IEEE Symposium on Security and Privacy (SP), pp. 121–140. IEEE

[41] Yoo, K., Ahn, W., Jang, J., Kwak, N.: Robust multi-bit natural language watermarking through invariant features. In: Proceedings of the 61st Annual Meeting

of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2092–2115

[42] Yang, X., Chen, K., Zhang, W., Liu, C., Qi, Y., Zhang, J., Fang, H., Yu, N.: Watermarking text generated by black-box language models. arXiv Preprint arXiv:2305.08883 (2023)

[43] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T.: A watermark for large language models. arXiv Preprint arXiv:2301.10226 (2023)

[44] Wang, L., Yang, W., Chen, D., Zhou, H., Lin, Y., Meng, F., Zhou, J., Sun, X.: Towards codable text watermarking for large language models. arXiv Preprint arXiv:2307.15992 (2023)

[45] Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S.: Can AI-generated text be reliably detected? arXiv Preprint arXiv:2303.11156 (2023)

[46] Liu, Y., Tang, S., Liu, R., Zhang, L., Ma, Z.: Secure and robust digital image watermarking scheme using logistic and RSA encryption. Expert Systems with Applications **97**, 95–105 (2018)

[47] Zhong, X., Das, A., Alrasheedi, F., Tanvir, A.: A brief, in-depth survey of deep learning-based image watermarking. Applied Sciences **13**(21), 11852 (2023)

[48] Passos, L.A., Jodas, D., Costa, K.A., Souza Júnior, L.A., Rodrigues, D., Del Ser, J., Camacho, D., Papa, J.P.: A review of deep learning-based approaches for deepfake content detection. Expert Systems **41**(8), 13570 (2024)

[49] Wang, S.-Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8695–8704

[50] Ju, Y., Jia, S., Ke, L., Xue, H., Nagano, K., Lyu, S.: Fusing global and local features for generalized AI-synthesized image detection. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 3465–3469. IEEE

[51] Mandelli, S., Bonettini, N., Bestagini, P., Tubaro, S.: Detecting GAN-generated images by orthogonal training of multiple CNNs. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 3091–3095. IEEE

[52] OpenAI: New AI Classifier for Indicating AI-Written Text. Accessed: March 17, 2024 (2023). https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

[53] GPTZero: AI Detection Tool. Accessed: March 17, 2024 (2023). https://gptzero.me/

[54] Yu, X., Qi, Y., Chen, K., Chen, G., Yang, X., Zhu, P., Zhang, W., Yu, N.: GPT paternity test: GPT generated text detection with GPT genetic inheritance. arXiv Preprint arXiv:2305.12519 (2023)

[55] Hu, X., Chen, P.-Y., Ho, T.-Y.: Radar: Robust AI-text detection via adversarial learning. arXiv Preprint arXiv:2307.03838 (2023)

[56] Tian, Y., Chen, H., Wang, X., Bai, Z., Zhang, Q., Li, R., Xu, C., Wang, Y.: Multiscale positive-unlabeled detection of AI-generated texts. arXiv Preprint arXiv:2305.18149 (2023)

[57] Originality.AI: AI Content Detection and Plagiarism Checker. Accessed: March 17, 2024 (2023). https://originality.ai/

[58] Turnitin Solutions on AI Writing. Accessed: March 17, 2024 (2022). https://www.turnitin.com/solutions/topics/ai-writing/

[59] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., Zou, J.: GPT detectors are biased against non-native english writers. arXiv Preprint arXiv:2304.02819 (2023)

[60] Vanderbilt University: Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector. Accessed: March 17, 2024 (2023). https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/

[61] OpenAI: DALL·E 2 Pre-training Mitigations. Accessed: 2024-09-25 (2024). https://openai.com

[62] OpenAI: An Update on Our Safety & Security Practices. Accessed: 2024-09-25 (2024). https://openai.com/index/update-on-safety-and-security-practices

[63] Stability AI: Artificial Intelligence and the Content Ecosystem. Accessed: 2024-09-25 (2024). https://stability.ai/artificial-intelligence-and-the-content-ecosystem

[64] Google AI: Google AI Model Documentation and Responsible AI Practices. Accessed: 2024-09-25 (2024). https://ai.google/discover/palm2

[65] Chen, L., Zaharia, M., Zou, J.: How is ChatGPT's behavior changing over time? arXiv Preprint arXiv:2307.09009 (2023)

[66] Sanusi, T.: What Happens When AI Eats Itself. Accessed: June 13, 2024 (2023). https://deepgram.com/learn/when-ai-eats-itself

[67] Cyberspace Administration of China: Notice on the Issuance of the 2023 Regulations on Internet Security. Accessed: March 17, 2024 (2023). http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

[68] Madiega, T.: Artificial Intelligence Act. The AI Act, the first binding worldwide horizontal regulation on AI, sets a common framework for the use and supply of AI systems in the EU. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

[69] Parliament, E., European Union, C.: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending various regulations and directives (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, 12 July 2024. In force (2024). http://data.europa.eu/eli/reg/2024/1689/oj

[70] National Institute of Standards and Technology (NIST): Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Published in the Federal Register on November 1, 2023. (2023). https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[71] Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., Prado, M.L., Herrera-Viedma, E., Herrera, F.: Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. Information Fusion **99**, 101896 (2023)

[72] Knott, A., Pedreschi, D., Chatila, R., Chakraborti, T., Leavy, S., Baeza-Yates, R., Eyers, D., Trotman, A., Teal, P.D., Biecek, P.: Generative AI models should include detection mechanisms as a condition for public release. Ethics and Information Technology **25**(4), 55 (2023)