E-ISSN: 2997-9439

SemantJournals

American Journal of Education and Evaluation Studies
https://semantjournals.org/index.php/ AJEES

| Research Article

Check for updates

Things Get Strange When AI Starts Training Itself and its How it Should be Regulated

Mardonov Amirzhon Sherzod ugli
Tashkent State University of Law, Lecturer of Cyber Law Department
amirmardonov39@gmail.com

Abstract: The article explores the evolving role of artificial intelligence (AI) as it transitions into self-training systems capable of recursive self-improvement. It begins by showcasing how AI enhances industries like healthcare, finance, and entertainment, demonstrating its integration into modern life. The narrative then shifts to self-training AI, explaining its ability to learn autonomously, optimize performance, and adapt without direct human input. Key concepts like data utilization, pretext tasks, feature learning, and fine-tuning are discussed, illustrating how these mechanisms empower AI to handle complex tasks with minimal manual intervention.

Keywords: self-training ai, recursive self-improvement, data utilization, pretext tasks, fine-tuning, model autophagy disorder (MAD), ethical ai, ai regulation and data governance.

Introduction

In the 21st century, artificial intelligence (AI) has emerged as a transformative force, reshaping industries and daily life in ways that were once unimaginable. From simplifying routine tasks to tackling complex decisions, AI has seamlessly integrated into countless aspects of human activity, boosting efficiency, convenience, and productivity. For instance, in healthcare, AI-powered tools like IBM Watson help doctors diagnose diseases by analyzing vast amounts of medical data, suggesting treatments, and improving patient care outcomes[1].

In finance, AI-driven algorithms predict stock market trends and execute trades at speeds far beyond human capabilities, giving investors a competitive edge. Entertainment platforms such as Netflix and Spotify also rely on AI to personalize user experiences by recommending movies, shows, or music based on individual preferences. These examples illustrate how AI has woven itself into the fabric of modern life, offering solutions that were once the stuff of science fiction.

However, AI has entered a new era: self-improvement. Self-improving AI, often referred to as recursive self-improvement, describes an artificial intelligence system that is capable of improving its own capabilities through a cycle of learning and modification. Self-training of AI is

[1] See What is artificial intelligence in medicine? https://www.ibm.com/topics/artificial-intelligence-medicine.

capability enables systems to analyze their own processes, learn from mistakes, and autonomously optimize their performance[2].

This evolution is powered by advancements in machine learning and computational power, marking a profound shift in how AI interacts with the world.

For example:

- Google DeepMind's AlphaGo learned strategies that surpassed human capabilities in Go, refining its gameplay autonomously with each iteration.
- Tesla's Autopilot continuously processes real-world driving data, allowing it to improve decision-making without direct intervention. The car uses "computer vision" and an artificial intelligence (AI) technology called end-to-end machine learning that translates the images into driving decisions "With this launch, I believe we are close to the 'ChatGPT moment' for robotaxis, when they will truly take off and become part of our mainstream transport landscape," Crijn Bouman, co-founder and CEO of Rocsys, a hands-free electric vehicle charging solution company based in the Netherlands, told Euronews Next. Tesla is not the only company to roll out autonomous driving technology[3].

While the promise of autonomous AI is immense, it also raises critical concerns around ethics, safety, and transparency. For self-improving AI to thrive responsibly, it must align with human values and operate within robust frameworks designed to prevent misuse or unintended consequences. These challenges underline the need for careful oversight and regulation to ensure that the benefits of this technology can be fully realized without compromising trust or safety.

How AI Self-Training Works

AI self-training is a fascinating process that allows models to learn and improve without relying on manually labeled data, making it highly efficient and adaptable.

The first step in this process, data utilization, involves analyzing vast amounts of raw, unlabeled data to uncover patterns and structures. Unlike traditional AI systems that depend on explicit labels (e.g., "this is a cat, this is a dog"), self-training AI identifies relationships within the data itself, grouping similar items based on shared features. For example, it might cluster photos of cats together without being explicitly told that they depict cats. This method enables AI to process and make sense of data in ways that are both scalable and resource-efficient.

Another crucial aspect of AI self-training is the use of pretext tasks, which are auxiliary problems designed to help the model develop a foundational understanding of data without requiring labeled datasets. These tasks are intentionally simple and self-contained, yet they play a significant role in equipping AI models with the ability to recognize patterns and extract meaningful insights. For example, a text-based AI might be tasked with predicting missing words in a sentence, such as filling in the blank for "The ___ is blue." This process helps the AI learn the nuances of grammar, syntax, and context, making it better equipped for more complex language tasks.

Another essential element of AI self-training is feature learning, a process where AI models focus on identifying the most critical and informative aspects of data to solve specific problems. Instead of treating all data equally, feature learning enables the model to recognize patterns and correlations that are most relevant to the task at hand. This allows the AI to prioritize meaningful

[2] See Archit Sharma, Ahmed M. Ahmed, Rehaan Ahmad, "Chelsea Finn Self-Improving Robots: End-to-End Autonomous Visuomotor Reinforcement Learning" https://arxiv.org/abs/2303.01488?utm_source=chatgpt.com.

[3] See Euronews Tesla's Cybercab autonomous vehicle revealed: Is this the 'ChatGPT moment for robotaxis'?, https://www.euronews.com/next/2024/10/11/teslas-cybercab-autonomous-vehicle-revealed-is-this-the-chatgpt-moment-for-robotaxis

data while filtering out irrelevant or redundant information, ultimately improving its accuracy and efficiency.

Fine-tuning is a critical step in AI training that allows a model to move from general knowledge to a specialized understanding of a specific task or domain. While the model initially learns broad patterns from large datasets, fine-tuning adjusts its focus using a smaller, labeled dataset tailored to a particular application. This process essentially refines the model's capabilities, enabling it to excel in highly specific areas while building on its foundational knowledge. For example, OpenAI's Codex, which is trained on a wide array of programming languages, becomes even more effective when fine-tuned for tasks like generating Python code for data analysis or writing SQL queries for database management. By honing in on task-specific nuances, fine-tuning transforms a general-purpose model into a highly specialized tool[4].

Analogy: Buttering Bread

To simplify, think of AI as a slice of bread:

- Initial training places a pat of butter in the center (basic knowledge).
- Self-training spreads the butter evenly across the bread (refining knowledge to cover all areas effectively).

This process ensures the bread (AI model) is more uniformly coated (better at solving problems), even though no new butter (entirely new data) has been added[5].

Types of Self-Training AI and Their Uses

AI self-training methodologies vary in how they learn from data. Below is a detailed explanation of three main types—Supervised Learning, Unsupervised Learning, and Reinforcement Learning—along with real-world applications and in-depth examples.

1. Supervised Learning

What It Is: Supervised learning involves training a model using labeled datasets, where each input has a corresponding output. The AI learns the relationship between the input and output, allowing it to make predictions or classifications.

How It Works:

- Data consists of pairs: input (e.g., transaction details) and expected output (e.g., "fraudulent" or "legitimate").
- The model learns by minimizing the error between its predictions and the actual labels during training.

Real-World Example: Fraud Detection:

- Financial Institutions: Banks use supervised learning to detect fraudulent transactions. Models are trained on historical data containing both legitimate and fraudulent transactions.

2. Unsupervised Learning

What It Is: Unsupervised learning deals with unlabeled datasets. The model identifies patterns, structures, or relationships within the data without explicit guidance.

[4] See Self-Supervised Learning Harnesses the Power of Unlabeled Data, https://shelf.io/blog/self-supervised-learning-harnesses-the-power-of-unlabeled-data/?utm_source=chatgpt.com

[5] Things Get Strange When AI Starts Training Itself What happens if AI becomes even less intelligible? By Matteo Wong, https://www.theatlantic.com/technology/archive/2024/02/artificial-intelligence-self-learning/677484/

How It Works:

- The model groups data points into clusters or reduces data dimensions to find meaningful patterns.
- Algorithms include clustering (e.g., K-means) and dimensionality reduction (e.g., PCA).
- Real-World Example: Spotify Recommendations:
- Clustering User Preferences: Spotify groups users based on their listening habits. For example, users who frequently listen to indie rock are clustered together, and recommendations are made based on the preferences of others in the same cluster.

3. Reinforcement Learning

What It Is: Reinforcement learning trains an agent to take actions in an environment to maximize cumulative rewards. It learns through trial and error, exploring and exploiting its environment.

How It Works:

- An agent takes actions in an environment and receives feedback in the form of rewards or penalties.
- Over time, the agent develops strategies to maximize rewards while minimizing penalties.
- Real-World Example: AlphaZero by Google DeepMind:
- Mastering Chess and Go: AlphaZero learned to play chess and Go by playing millions of games against itself. Unlike traditional programs that rely on pre-programmed strategies, AlphaZero started with the basic rules and developed innovative strategies by learning from its own successes and failures.\

These types clearly show the difference between each other and it seems Reinforcement learning is becoming more popular un LLMs, because of its more efficient results[6].

Advantages of Self-Training AI

Self-training AI offers several advantages that make it a game-changer across industries. Its adaptability allows models to update themselves with new data, such as Netflix refining recommendations as viewers explore new genres. Data efficiency reduces dependency on large datasets; tools like AlphaFold predict protein structures with minimal experimental data, accelerating drug discovery. In cybersecurity, platforms like Darktrace monitor and mitigate threats in real-time, while autonomous vehicles continuously improve their safety and performance through real-world feedback. Moreover, Self-training system of AI in bank cybersecurity systems or governmental systems are becoming very popular now, by this way they can prevent many attacks that might happen. These benefits highlight how self-training AI can transform operations, enhance decision-making, and drive innovation.

Strange and Unexpected Outcomes

Despite its benefits, self-training AI can produce unexpected results. For instance, bias amplification can occur, as seen in Amazon's hiring AI, which favored male candidates due to biased training data. Similarly, models trained on noisy data may overfit, leading to errors such as wildlife monitoring systems mistaking animals based on irrelevant features like timestamps. Language models like ChatGPT occasionally generate contextually irrelevant responses, and image classification models sometimes generalize incorrectly, mistaking cats for dogs based on

[6] See Pragya Soni - Self Supervised Learning - Types, Examples and Applications, https://www.analyticssteps.com/blogs/self-supervised-learning-types-examples-and-applications?utm_source=chatgpt.com

shared background elements. These challenges underscore the importance of robust training data and oversight. By upcoming scientific result, it will be clear how it might cause unexpected consequences.

Lessons from Fiction: Self-Training in Pop Culture

When we are talking about self-training of AI – we can remember a few movies or cartoons here, but **Cloudy with a chance of meatballs** that might use some aspect of this topic, the protagonist Flint Lockwood creates the FLDSMDFR (Food Lengthening and Storage Mechanism Designated Food Replicator), a machine designed to convert water into food. This machine can be viewed as an initial AI system—designed with specific instructions or "codes" to perform a task. These instructions are similar to the data fed into a machine learning model in the real world.

In the context of AI, adding "codes as data" means feeding the system structured input to train it. This could be the initial training dataset for an AI, where the system learns from provided data and is expected to generate results based on it. In the case of Flint’s machine, this is when the FLDSMDFR begins functioning as intended—producing food from water.

Process of the self-training of the machine:

- ✓ *Adding codes as data;*
- ✓ *Achieving an expected result;*
- ✓ *Improving the result by self-training;*
- ✓ *Overimroving its capacity which led to bad consequences.*

This story serves as a metaphor for the broader concerns around AI self-training and optimization: the need for constant monitoring, control mechanisms, and alignment with human priorities to avoid negative consequences.

Proof of these circumstances

Recently, Machine learning researchers Sina Alemohammad and Josue Casco-Rodriguez, both PhD students in Rice University's Electrical and Computer Engineering department, and their supervising professor, Richard G. Baraniuk. In collaboration with researchers at Stanford, they recently published a fascinating — though yet to be peer-reviewed — paper on the subject, titled "Self-Consuming Generative Models Go MAD." MAD, which stands for Model Autophagy Disorder, is the term that they've coined for AI's apparent self-allergy. In their research, it took only five cycles of training on synthetic data (are artificially generated data rather than produced by real-world events) for an AI model's outputs to, in the words of Baraniuk, "blow up[7]."

The unpredictability of syntactic data in AI reinforces the importance of careful model design, quality training data, and effective supervision to ensure that AI outputs remain reliable and coherent.

Regulating Self-Training AI in the world

Global efforts to regulate AI are underway to address risks associated with self-training systems.

As self-training AI becomes more advanced, it’s clear that proper regulation is essential to balance its benefits with its risks. Around the world, governments and organizations are creating rules to make sure this technology is used responsibly, especially in sensitive areas like healthcare, finance, and public services. While different countries have their own approaches, the overall goal is the same: to ensure AI is fair, safe, and trustworthy while protecting people’s rights.

[7] See When AI Is Trained on AI-Generated Data, Strange Things Start to Happen, https://futurism.com/ai-trained-ai-generated-data-interview

In **the European Union (EU)**, regulation is already very detailed. The AI Act is a new law that sorts AI systems into different risk levels. High-risk uses, like AI in hospitals or policing, have strict rules to make sure they are safe, transparent, and regularly checked. At the same time, the EU's General Data Protection Regulation (GDPR) ensures that personal data is handled with care, giving people more control over how their information is used. Together, these laws set a strong example of how to manage AI while protecting individual rights. The European Union's Artificial Intelligence Act (AI Act), effective as of August 1, 2024, establishes a comprehensive framework for AI regulation, categorizing AI systems based on risk levels and imposing corresponding requirements. However, the Act does not explicitly address self-training AI systems, leaving a regulatory gap in this rapidly evolving area[8].

The United States takes a more varied approach, with both state-level and federal efforts. For example, California's Consumer Privacy Act (CCPA) gives people the right to control their data—letting them ask companies to delete or stop collecting it. On a national level, the proposed Algorithmic Accountability Act aims to stop AI systems from being biased. This law would require companies to study how their AI systems affect people, ensuring they don't unfairly discriminate. However, since there isn't a single nationwide policy yet, AI regulation in the U.S. can feel inconsistent depending on where you are. See The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment by Alex Engler Tuesday, April 25, 2023[9]

In **Canada**, the AI and Data Act focuses on accountability and transparency. This proposed law would make companies responsible for how their AI systems behave, especially when those systems are used in public services. By emphasizing clear rules and ethical use, Canada hopes to build trust in AI while encouraging responsible innovation.

China has a different approach, tying its AI rules to broader government goals. Its guidelines stress ethical AI use and accountability, but they also focus on giving the government strong control over how AI is developed and deployed. While this allows China to quickly advance its AI technology, it raises questions about privacy and individual freedoms, especially with the rise of AI in surveillance and social credit systems.

In **Australia**, the AI Ethics Framework is a simpler, voluntary set of guidelines. It encourages developers to be transparent and fair, focusing on using AI for the greater good. While it's not legally binding, the framework lays the groundwork for future laws and ensures that ethical AI practices are part of the national conversation.

Despite these differences, only the U.S. has a unique approach here – the government of the USA has been working in cooperation with private sector and software developers and prevent the possible mistakes by AI and its self-training systems. This extraordinary research might be very efficient in the future because of the strong relationship between developers and bodies, thus some of the countries in our list are willing to integrate this strategy to their law systems. These developments indicate that the U.S. is proactively formulating policies to address the complexities of self-developing AI, potentially positioning itself as a leader in this regulatory domain.

Proposal for Comprehensive AI Regulation

Self-learning AI is changing how technology works, offering many benefits but also bringing risks that need to be managed carefully. Unlike traditional AI, which follows fixed rules or relies on human guidance, self-learning AI can adapt and improve on its own. This makes it more powerful but also more unpredictable. To keep this technology safe and fair, clear rules must be created. These rules should focus on defining self-learning AI, improving how data is handled,

[8] https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en?utm_source=chatgpt.com

[9] https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/?utm_source=chatgpt.com

making systems transparent, ensuring accountability, encouraging ethical practices, and monitoring AI over time.

The first step in regulating self-learning AI is to **define it clearly**. Traditional AI depends on human input and fixed programming, while self-learning AI has the ability to change and improve by learning from data. This distinction is important because self-learning AI can behave in unexpected ways if not managed properly. For example, a chatbot with self-learning capabilities might develop harmful or biased responses over time without anyone realizing it. A clear definition helps everyone understand how self-learning AI is different and why it needs specific regulations.

Another key part of regulation is **data governance**. Self-learning AI relies heavily on data to learn and improve, but if the data is biased or flawed, the AI will inherit those problems. For instance, an AI system used in hiring might favor certain groups over others if the training data is biased. To prevent this, companies should be required to use high-quality, diverse datasets and have their data audited regularly. Audits would help identify and fix issues early, ensuring the AI produces fair and accurate results.

Transparency is also essential for building trust in self-learning AI. Developers should explain how their AI systems work, including the data they use, how the AI is trained, and the reasons behind its decisions. For example, a bank using AI to approve loans should give clear explanations to customers about why their loan applications were accepted or denied. Transparency helps people trust AI systems and gives regulators the information they need to ensure these systems are safe and fair.

Accountability is equally important. If an AI system causes harm, such as a self-driving car getting into an accident, it should be clear who is responsible—whether it’s the company that built the car, the software developer, or the operator. Having clear accountability rules ensures that companies take safety seriously and that victims can receive compensation if something goes wrong. This also encourages developers to be more careful and ethical when creating AI systems.

To make sure AI systems are developed responsibly, **ethical oversight** within organizations is crucial. Companies should have internal teams or committees to review their AI projects and ensure they align with ethical standards. For example, if an AI system used in hiring is found to exclude certain groups unfairly, the ethics team could recommend changes to fix this. These committees play a key role in preventing problems and promoting fairness and inclusivity in AI systems.

Finally, **continuous monitoring and enforcement by government agencies** is necessary to ensure AI remains safe and effective over time. These agencies could test self-learning AI regularly to check for errors, bias, or unexpected behaviors. For example, self-driving cars could be tested under different road conditions to make sure they adapt properly without risking safety. Regular monitoring helps catch problems early and ensures AI systems continue to meet safety and ethical standards. See Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities.

Why Self-Learning AI Matters for lawyers

Lawyers should care about AI training itself because it raises significant ethical and legal implications. As AI systems become more autonomous, issues such as accountability, bias, and data privacy emerge. Lawyers need to understand these challenges to effectively navigate potential liabilities and ensure compliance with legal standards, ultimately safeguarding client interests and maintaining the integrity of the legal profession. By addressing these concerns, self-training AI can evolve as a force for good, transforming industries while safeguarding human interests.

Conclusion

Self-training AI represents a revolutionary leap in technology, with the potential to transform industries and redefine the way we approach complex problems. From supervised learning in fraud detection and medical diagnostics to unsupervised learning in personalized recommendations and reinforcement learning in autonomous vehicles, the versatility and adaptability of AI systems are unparalleled. These methods, combined with hybrid approaches like semi-supervised and self-supervised learning, have expanded the boundaries of AI capabilities, making it more efficient and scalable while reducing dependency on manual data labeling.

However, as this technology evolves, it brings with it challenges that demand careful attention. Bias amplification, overfitting, and unexpected outcomes highlight the critical need for robust training data, ethical oversight, and constant monitoring. Fictional analogies, like Flint Lockwood’s machine in *Cloudy with a Chance of Meatballs*, serve as cautionary tales, reminding us of the dangers of unchecked self-improvement in AI. Real-world examples, such as the "Model Autophagy Disorder" observed by researchers, further emphasize the importance of thoughtful design and regulation.

Regulatory efforts across the globe, from the EU’s AI Act to the U.S. Algorithmic Accountability Act, are steps in the right direction. These frameworks aim to ensure that AI development is transparent, ethical, and aligned with societal values. For professionals, including lawyers, understanding self-training AI is essential to navigate its ethical, legal, and social implications. This awareness ensures that while AI continues to drive innovation and efficiency, it also remains a force for good, serving human interests responsibly.

As we look to the future, the promise of self-training AI lies not only in its ability to solve today’s problems but also in its capacity to anticipate and address challenges we have yet to encounter. By embracing the benefits of this technology while mitigating its risks, we can harness the transformative power of AI to create a better, more equitable world.

List of references:

1. https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/?utm_source=chatgpt.com
2. https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en?utm_source=chatgpt.com
3. Eshonkulov J. (2025). The Role of Smart Contracts in Civil Law and Issues of Legal Regulation. Uzbek Journal of Law and Digital Policy, 3(1), 104–111. https://doi.org/10.59022/ujldp.294
4. Eshonkulov, J. (2024). Legal foundations for the application of artificial intelligence Technologies in the Sports Industry. American Journal of Education and Evaluation Studies, 1(7), 240-247. https://semantjournals.org/index.php/AJEES/article/view/320/287
5. When AI Is Trained on AI-Generated Data, Strange Things Start to Happen, https://futurism.com/ai-trained-ai-generated-data-interview
6. Pragya Soni - Self Supervised Learning - Types, Examples and Applications, https://www.analyticssteps.com/blogs/self-supervised-learning-types-examples-and-applications?utm_source=chatgpt.com
7. Self-Supervised Learning Harnesses the Power of Unlabeled Data, https://shelf.io/blog/self-supervised-learning-harnesses-the-power-of-unlabeled-data/?utm_source=chatgpt.com
8. Things Get Strange When AI Starts Training Itself What happens if AI becomes even less

intelligible? By Matteo Wong, https://www.theatlantic.com/technology/archive/2024/02/artificial-intelligence-self-learning/677484/

9. Euronews Tesla's Cybercab autonomous vehicle revealed: Is this the 'ChatGPT moment for robotaxis'?, https://www.euronews.com/next/2024/10/11/teslas-cybercab-autonomous-vehicle-revealed-is-this-the-chatgpt-moment-for-robotaxis
10. What is artificial intelligence in medicine? https://www.ibm.com/topics/artificial-intelligence-medicine
11. Archit Sharma, Ahmed M. Ahmed, Rehaan Ahmad, “Chelsea Finn Self-Improving Robots: End-to-End Autonomous Visuomotor Reinforcement Learning” https://arxiv.org/abs/2303.01488?utm_source=chatgpt.com