Article

Not peer-reviewed version

Non-Obvious Alien AI Constructions: Opportunities and Implications

Sean Khozin *

Posted Date: 27 August 2024

doi: 10.20944/preprints202408.1929.v1

Keywords: artificial intelligence; machine learning; biomdical research; drug dicovery and development

 doi:10.20944/preprints202408.1929.v1

Article

Non-Obvious Alien AI Constructions: Opportunities and Implications

Sean Khozin

Massachusetts Institute of Technology, Cambridge, MA, United States of America | CEO Roundtable on Cancer and Project Data Sphere, Morrisville, NC | Phyusion Bio, LLC, New York, NY; khozin@mit.edu

Abstract: This paper examines the utility and exploitation of non-obvious "alien" AI constructions, artificial intelligence solutions that significantly deviate from human intuition and traditional problem-solving approaches. It explores recent revelations, key characteristics, applications across various domains, and far-reaching implications for scientific research and innovation. Particular attention is given to developments in mathematical reasoning, drug discovery, and healthcare. The paper also discusses challenges in implementation, ethical considerations, and future directions, including meta-learning algorithms and human-AI collaboration frameworks. By synthesizing current research and identifying emerging trends, this paper aims to provide a basic understanding of the transformative potential of non-obvious AI solutions in reshaping scientific discovery and biomedical innovation.

Keywords: artificial intelligence; machine learning; biomdical research; drug dicovery and development

---

1. Introduction

The concept of non-obvious AI constructions is a largely overlooked theme in artificial intelligence (AI) and problem-solving. Non-obvious AI solutions are characterized by their radical departure from traditional human thinking patterns, often appearing counterintuitive or even incomprehensible to human experts at first glance. The term "alien" in this context refers to the foreign and unexpected nature of these AI-generated approaches, which can transcend human cognitive biases and explore previously overlooked solution spaces.

The origins of this concept can be traced back to landmark events in AI development. A pivotal moment occurred in 2016 during the match between DeepMind's AlphaGo and Go grandmaster Lee Sedol. AlphaGo's now-famous "Move 37" in the second game demonstrated the AI's capacity to make plays that defied conventional Go strategy, accumulated over millennia of human expertise [1]. This event signaled AI's potential to transcend human cognitive limitations and explore solution spaces previously deemed implausible or entirely overlooked.

Definition 1 (Non-obvious Alien AI Construction. Please note that in this paper **non-obvious** and **alien** are used interchangeably)**.** *Let S be the set of all possible solutions and $H \subset S$ be the set of human-conceivable solutions. A solution $s^* \in S$ is a non-obvious alien AI construction if:*

1. *s^* is optimal: $s^* \in argmax_{s \in S} h(s)$, where $h : S \to R$ measures solution quality.*
2. *s^* is not human-conceivable: $s^* \notin H$.*
3. *s^* is significantly different from human solutions: $\min_{s \in H} d(s, s^*) > \epsilon$, for some $\epsilon > 0$.*
4. *s^* is unlikely to be conceived by humans: $P(s^* \mid \text{human knowledge}) < \delta$, for small $\delta > 0$.*

2. Recent Revelations in Alien AI Constructions

2.1. Mathematical Reasoning

In July 2024, Google DeepMind introduced an AI system that demonstrated exceptional capabilities in mathematical problem-solving, achieving near-gold medal performance at the International Mathematical Olympiad (IMO) [2]. This system, which combines Alpha Proof and Alpha Geometry 2, successfully tackled complex problems, including ones that were attempted by only a handful of human contestants. As Professor Sir Timothy Gowers, a Fields Medalist, observed, the AI's capacity to

 doi:10.20944/preprints202408.1929.v1

generate its solutions showcases a level of reasoning and creativity that appears almost alien, even to experienced mathematicians.

The concept of non-obvious alien constructions underscores the utility of AI to transcend human cognitive limitations and explore solution spaces that might be overlooked or deemed implausible by human experts. This capability allows AI systems to approach problems from completely novel perspectives.

The integration of AI systems like Alpha Proof and Alpha Geometry 2 into research methodologies enables the exploration of these unconventional solutions. Such systems can be thought of as not just computational tools but also sources of new hypotheses and problem-solving strategies that in the short-term can complement and extend human expertise. This collaborative dynamic between AI and human researchers fosters opportunities in complex problem-solving, where AI-driven insights challenge existing theories and open up new avenues for inquiry.

2.2. Drug Discovery and Development

In the pharmaceutical industry, non-obvious AI solutions can significantly improve drug discovery and development processes. Traditional approaches often involve iterative refinements of known molecular structures or screening vast compound libraries. However, AI systems capable of generating non-obvious solutions can enable the development of *de novo* AI-powered drug designs [3].

For instance, Segler et al. demonstrated an AI system that could generate focused libraries of novel, drug-like molecules [4]. The system used recurrent neural networks trained on large datasets of known molecules to propose new structures. Importantly, many of these AI-generated molecules were not obvious derivatives of existing drugs, but rather novel constructions that human chemists might not have considered.

Furthermore, non-obvious AI solutions are being applied to other aspects of drug development, such as predicting drug-target interactions, optimizing lead compounds, and designing more efficient clinical trials [5]. These AI-driven approaches have the potential to significantly reduce the time and cost associated with bringing new drugs to market and are poised to fundamentally reshape traditional approaches to drug discovery and development.

2.3. Healthcare and Precision Medicine

Non-obvious AI solutions are critical in advancing precision medicine and healthcare delivery goals, particularly in the area of individualized treatments. Rather than relying on broad categories and averaged responses, AI's ability to identify subtle, non-obvious patterns in patient data allows for the identification of previously unrecognized patient subgroups or biomarkers [6].

For example, a study by Rajkomar et al. used deep learning models to analyze electronic health records and predict a range of clinical outcomes [7]. The models identified non-obvious predictors of outcomes that were not part of traditional clinical scoring systems, demonstrating AI's potential to uncover hidden patterns in complex medical data.

This approach is the foundation of enabling more precise and personalized medical interventions. By analyzing complex patient data, including genomic, proteomic, and clinical information, AI can identify subtle subgroups of patients who may respond differently to treatments [8]. This is particularly promising in oncology, where AI-driven analyses have revealed novel cancer subtypes and potential treatment strategies [9].

3. Characteristics of Non-Obvious Alien AI Constructions

Non-obvious AI solutions exhibit several key characteristics that distinguish them from conventional problem-solving approaches:

 doi:10.20944/preprints202408.1929.v1

3.1. Counterintuitive Approach

These solutions often tackle problems from angles that human experts might not consider, leading to unexpected results. This unconventional methodology allows AI systems to explore solution spaces that may be overlooked by traditional methods.

3.2. Enhanced Pattern Recognition

Modern AI architectures excel in identifying subtle patterns in vast datasets that may be imperceptible to human analysts. This enhanced "alien" pattern recognition enables AI to uncover insights and correlations that could remain hidden when using conventional analytical techniques.

4. Applications of Non-Obvious Alien AI Constructions Beyond Life Sciences

4.1. Materials Science

In materials science, non-obvious AI solutions can accelerate the discovery of new materials with desired properties. For example, Tshitoyan et al. demonstrated that an AI system trained on scientific literature could predict new thermoelectric materials years before their discovery by humans, highlighting the utility of their embedding approach in capturing complex materials science concepts beyond the confines of human cognitive capabilities [10].

4.2. Climate Science

Alien AI solutions are also surfacing in climate science. For instance, Reichstein et al. showed how AI can be used to improve climate models by identifying complex patterns in Earth system data that are difficult for humans to perceive [11].

4.3. Robotics

In robotics, Alien AI solutions can enable the development of more adaptable and efficient systems. For example, Hwangbo et al. demonstrated an AI system that could learn to control a quadrupedal robot in complex environments through a process of trial and error independent of human oversight [12].

5. Exploiting Alien AI Constructions through Latent Space Analysis

Recent research points to the potential of new methodologies for exploring and exploiting Alien solutions, particularly through the rigorous investigation of latent spaces in deep learning models. This section examines strategies for leveraging these hidden representations to uncover novel and potentially counterintuitive solutions to complex problems.

5.1. Latent Space Interpolation and Extrapolation

One promising approach involves the systematic exploration of the latent space through interpolation and extrapolation techniques. Jahanian et al. (2020) demonstrated that linear interpolation in the latent space of generative adversarial networks (GANs) can reveal semantically meaningful transformations that are not explicitly encoded in the training data [13]. Their work elucidates the "steerability" of GANs, showing that certain directions in the latent space correspond to interpretable image transformations.

Extending this concept to extrapolation, one can potentially generate solutions that lie beyond the boundaries of the training distribution. This extrapolation in latent space can be formalized as:

$$z_{extrapolated} = z_{base} + \alpha(z_{target} - z_{base}) \tag{1}$$

where z_{base} and z_{target} are latent vectors, and $\alpha > 1$ for extrapolation. This technique can lead to the discovery of non-obvious alien alternatives by exploring regions of the latent space that are outside the convex hull of the training data.

 doi:10.20944/preprints202408.1929.v1

5.2. Adversarial Latent Space Manipulation

Leveraging adversarial techniques to manipulate the latent space offers another avenue for uncovering non-obvious solutions. Shen et al. (2020) proposed a method called InterFaceGAN, which identifies interpretable directions in the latent space of pre-trained GANs [14]. Their approach involves training a binary classifier to separate latent codes of images with and without a specific attribute, then using the normal vector to the decision boundary as the editing direction.

Formally, given a latent code z and an editing direction n, the edited latent code is:

$$z_{edited} = z + \alpha n \tag{2}$$

where α controls the degree of manipulation. This method allows for fine-grained control over specific attributes while maintaining overall image coherence, potentially revealing non-obvious combinations of features.

5.3. Disentangled Representation Learning

Disentangled representation learning aims to separate the underlying factors of variation in the data within the latent space. Locatello et al. (2021) introduced a framework for weakly-supervised disentanglement that leverages limited supervision to achieve more interpretable and controllable latent representations [15]. Their method, based on the principle of independent mechanisms, uses a small number of observations of groups of correlated attributes to learn disentangled representations.

The authors propose a modification to the β-VAE objective:

$$\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta KL(q_\phi(z|x)||p(z)) + \gamma \mathcal{L}_{weak} \tag{3}$$

where $\mathcal{L}_{weak}$ is a weakly-supervised loss term that encourages disentanglement based on the available grouped observations. This approach facilitates the exploration of latent space dimensions that correspond to semantically meaningful factors, potentially leading to the discovery of non-obvious solutions through targeted manipulation of specific attributes.

5.4. Topological Data Analysis of Latent Spaces

Applying topological data analysis (TDA) to latent spaces can be a viable approach to uncovering alien complex structures and relationships that may not be apparent through traditional analytical methods [16]. TDA can be used in analyzing the latent space of variational autoencoders (VAEs) to reveal hidden patterns in, for example, single-cell RNA sequencing data. Their topological autoencoder incorporates persistent homology into the VAE framework, ensuring that the learned representations preserve important topological features of the data.

The topological loss term in their model is defined as:

$$\mathcal{L}_{topo} = \sum_{k=0}^{K} W_k d_B(\mathrm{PD}_k(X), \mathrm{PD}_k(Z)) \tag{4}$$

where $\mathrm{PD}_k(X)$ and $\mathrm{PD}_k(Z)$ are the k-dimensional persistence diagrams of the input and latent space respectively, and d_B is the bottleneck distance between persistence diagrams. This approach can be extended to other domains, potentially revealing non-obvious solutions embedded in the topological structure of the latent space.

5.5. Multi-Modal Latent Space Fusion

Integrating latent representations from multiple modalities can lead to the discovery of aliem solutions that leverage complementary information across different data types. Shi et al. (2020) proposed a contrastive learning framework for multi-modal fusion that aligns latent spaces from different modalities while preserving modality-specific information [17]. Their approach, applied to unpaired

 doi:10.20944/preprints202408.1929.v1

image-to-image translation, uses a contrastive loss to encourage the model to map corresponding images from different domains close to each other in the latent space.

The contrastive loss is defined as:

$$\mathcal{L}_{contrast} = -\log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{k=1}^{N} \exp(z_i \cdot z_k/\tau)} \tag{5}$$

where z_i and z_j are latent representations of corresponding images from different domains, τ is a temperature parameter, and N is the number of negative samples. This method enables the exploration of solutions that may not be apparent when considering each modality in isolation, potentially uncovering non-obvious relationships between different data types.

5.6. Evolutionary Algorithms in Latent Space

Applying evolutionary algorithms to latent space representations can offer a powerful method for discovering alien constructions through iterative optimization. Gaier and Ha (2019) introduced a technique called Weight Agnostic Neural Networks, which uses a quality diversity algorithm to explore the latent space of network architectures, producing diverse and high-performing solutions [18]. Their approach combines the generative power of deep learning with the exploratory capabilities of evolutionary algorithms.

The fitness function for their evolutionary algorithm is defined as:

$$f(a) = \sum_{w \in W} \text{performance}(a, w) + \lambda \cdot \text{diversity}(a) \tag{6}$$

where a is a network architecture, W is a set of shared weight values, and λ is a hyperparameter controlling the trade-off between performance and diversity. This method can potentially uncover solutions that neither deep learning nor evolutionary computation could find independently, by exploring the space of network architectures in a way that is agnostic to specific weight values.

5.7. Causal Structure Learning in Latent Space

Inferring causal relationships within the latent space can provide insights into the underlying mechanisms governing complex systems, potentially leading to the discovery of non-obvious interventions or solutions. Shen et al. (2022) proposed a method for learning causal structures in the latent space of deep generative models, enabling the identification of causal factors and their relationships [14]. Their approach, called Masked Latent Causal Learning, uses a masked self-attention mechanism to model causal dependencies between latent variables.

The causal structure is learned by optimizing the following objective:

$$\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)||p(z)) + \lambda \mathcal{L}_{causal} \tag{7}$$

where $\mathcal{L}_{causal}$ is a regularization term that encourages the learned latent representation to reflect the true causal structure of the data. This approach can be particularly valuable in domains where understanding causal mechanisms is crucial for developing effective solutions, as it allows for the identification of alien causal relationships that may not be apparent from observational data alone.

By leveraging these advanced techniques for latent space analysis, one can systematically explore and exploit the hidden potential of AI models, uncovering alien constructions that could significantly widen the aperture of problem-solving across various domains. As these methods continue to evolve, they promise to push the boundaries of possibility, revealing new approaches to addressing complex real-world challenges.

6. Challenges and Considerations

While the potential of alien AI solutions is immense, their implementation comes with several challenges.

 doi:10.20944/preprints202408.1929.v1

6.1. Interpretability

The "black box" nature of many AI-derived alien constructions can make it difficult to interpret or explain their decision-making processes. This lack of transparency can be problematic, especially in high-stakes decisions such as healthcare and biomedical research. While the field is working on developing interpretable machine learning models [19], there is an ongoing debate about the trade-off between interpretability and performance.

6.2. Ethical Considerations

As AI systems generate increasingly alien solutions, ensuring their alignment with human values becomes paramount. The potential for unintended consequences or misaligned objectives necessitates careful consideration of ethical implications and the development of robust governance frameworks [20].

6.3. Validation and Trust

Alien AI constructions may face skepticism from human experts, particularly when they contradict established knowledge or intuition. Developing rigorous validation methodologies and building trust in AI-generated solutions will be crucial for their widespread adoption [21].

6.4. Human-AI Collaboration Frameworks

Effective implementation of non-obvious AI solutions often requires integration with human expertise, at least until affected workflows adapt to autonomous decision-making. Therefore, striking the right balance between AI-driven innovation and human judgment remains a key challenge [22]. The integration of AI into human workflows and processes is rapidly evolving, offering new avenues for solving complex problems. A critical component in leveraging this potential is the development of effective human-AI collaboration frameworks. These frameworks are designed to enhance the synergy between human intuition and AI computational reasoning, ultimately leading to alien solutions that neither could achieve independently.

6.4.1. Interactive Visualization Tools

Interactive visualization tools are essential in bridging the gap between human users and AI systems. These tools transform complex data outputs from AI into comprehensible visual formats, allowing humans to better understand, interpret, and engage with AI-generated insights. For instance, tools that visualize decision paths in machine learning models can help users grasp why a particular decision was made, thereby increasing transparency and trust. Additionally, interactive dashboards enable users to manipulate data inputs and observe potential outcomes, fostering a deeper understanding of the AI's decision-making process and facilitating more informed decision-making.

6.4.2. Natural Language Interfaces

Natural Language Interfaces (NLIs) are another crucial element in human-AI collaboration frameworks. By allowing users to interact with AI systems using everyday language, NLIs lower the barrier to entry and make advanced AI capabilities accessible to non-experts. This can be particularly beneficial in domains where subject matter expertise is critical but technical expertise may be lacking. For example, healthcare professionals can use NLIs to query AI systems about patient data or treatment options, receiving explanations and recommendations in understandable terms. This natural mode of interaction not only enhances usability but also ensures that AI systems are used effectively and appropriately.

6.4.3. Collaborative Reasoning Systems

Collaborative reasoning systems aim to facilitate joint problem-solving between humans and AI. These systems enable a collaborative exchange of ideas, where AI provides data-driven insights and

 doi:10.20944/preprints202408.1929.v1

humans contribute contextual understanding and creative thinking. A key aspect of these systems is the ability to dynamically adjust the level of AI involvement based on the user's expertise and preferences. For example, in collaborative writing tools, AI agents can suggest edits or additions while the human author retains control over the final content, blending computational efficiency with human creativity.

6.4.4. Conventions and Adaptability in Human-AI Collaboration

Andy Shih et al. emphasize the importance of conventions in adaptive human-AI collaboration [23]. Just as humans develop shared conventions through repeated interactions—such as signals in sports or jargon in professional contexts—AI systems need to learn and adapt to these conventions to collaborate effectively. This involves distinguishing between rule-dependent behavior, which pertains to the fundamental rules of a task, and convention-dependent behavior, which arises from the specific practices and preferences of the human partner.

To facilitate seamless adaptation to new partners and tasks, AI systems can leverage a dual representation framework: one for task-specific rules and another for partner-specific conventions. By learning these representations separately, AI can quickly adapt to new partners by reusing the rule-based knowledge while adjusting to the new partner's conventions. This approach not only improves the efficiency of the collaboration but also enhances the AI system's ability to support a diverse range of users and scenarios.

7. Future Directions

7.1. Metalearning Methods

Metalearning, often described as "learning to learn," is a critical topic in the field of AI and machine learning. This approach aims to develop algorithms that can learn from their own learning experiences, thereby improving their performance and adaptability across various tasks and domains. These systems typically employ a two-tiered process of base learning and meta-learning, allowing rapid adaptation to new problems and novel solution generation in unfamiliar contexts.

Hospedales et al. (2022) provide a comprehensive survey of metalearning in neural networks, highlighting its potential to address some of the key challenges in current AI systems [24]. The authors define metalearning as a paradigm that "aims to learn algorithms that can learn efficiently on unseen tasks with a few examples."

Key aspects of metalearning include:

1. **Few-shot Learning:** Metalearning algorithms are designed to perform well on new tasks with very limited training data. This capability is crucial for developing AI systems that can quickly adapt to novel situations, mirroring human-like learning abilities that can evolve into alien constructions.
2. **Transfer Learning:** Metalearning facilitates better transfer of knowledge across different but related tasks. This allows AI systems to leverage previously acquired knowledge to solve new problems more efficiently.
3. **Hyperparameter Optimization:** Metalearning can automate the process of tuning model hyperparameters, a traditionally time-consuming and expertise-dependent task in machine learning.
4. **Architecture Search:** Some metalearning approaches can automatically discover optimal neural network architectures for specific tasks, potentially leading to more efficient and effective AI models.

The survey by Hospedales et al. identifies several key approaches to metalearning in neural networks:

- **Metric-based methods:** These focus on learning a metric space where similar examples are close together, facilitating few-shot learning.

 doi:10.20944/preprints202408.1929.v1

- **Model-based methods:** These approaches aim to design neural network architectures that are inherently quick to fine-tune with new information.
- **Optimization-based methods:** These methods learn update rules or initialization parameters that allow for rapid adaptation to new tasks.

In the context of non-obvious alien AI constructions, metalearning holds significant promise. By enabling AI systems to adapt more quickly and effectively to new problems, metalearning could lead to the discovery of novel solutions that might not be apparent through traditional learning approaches. For example, a metalearning system applied to drug discovery could potentially identify unconventional molecular structures or drug targets by rapidly adapting its learning strategy based on limited data from various chemical and biological domains. Similarly, in climate science, a metalearning approach could help in developing models that quickly adapt to new types of environmental data, potentially uncovering non-obvious patterns in climate change dynamics.

However, as Hospedales et al. point out, there are still challenges in the field of metalearning. These include the need for more theoretical understanding of why certain metalearning algorithms work, scalability issues with some approaches, and the challenge of designing truly general-purpose metalearning systems.

As research in this area progresses, we can expect metalearning to play an increasingly important role in developing AI systems capable of generating non-obvious, alien solutions to complex problems across various domains.

8. Conclusion

Non-obvious alien AI constructions represent a fundamental shift in our approach to problem-solving and innovation. By generating solutions that transcend human intuition and cognitive biases, these AI systems are opening up new frontiers in scientific research and technological development. While challenges remain, particularly in terms of interpretability, ethical considerations, and human-AI integration, the potential benefits of these alien constructions are immense. As we continue to explore and harness these capabilities, we may find that the future of scientific research and innovation is more unconventional and promising than we can currently imagine.

References

1. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; others. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489.
2. Google DeepMind. AI achieves silver-medal standard solving International Mathematical Olympiad problems, 2024. Press Release.
3. Schneider, P.; Walters, W.P.; Plowright, A.T.; Sieroka, N.; Listgarten, J.; Goodnow Jr, R.A.; Fisher, J.; Jansen, J.M.; Duca, J.S.; Rush, T.S.; others. Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery* **2020**, *19*, 353–364.
4. Segler, M.H.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science* **2018**, *4*, 120–131.
5. Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opinion on Drug Discovery* **2021**, *16*, 949–959.
6. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *New England Journal of Medicine* **2019**, *380*, 1347–1358.
7. Rajkomar, A.; Oren, E.; Chen, K.; Dai, A.M.; Hajaj, N.; Hardt, M.; Liu, P.J.; Liu, X.; Marcus, J.; Sun, M.; others. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* **2018**, *1*, 1–10.
8. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* **2019**, *25*, 44–56.
9. Wiens, J.; Shenoy, E.S. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases* **2018**, *66*, 149–153.

 doi:10.20944/preprints202408.1929.v1

10. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K.A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95–98.
11. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204.
12. Hwangbo, J.; Lee, J.; Dosovitskiy, A.; Bellicoso, D.; Tsounis, V.; Koltun, V.; Hutter, M. Learning agile and dynamic motor skills for legged robots. *Science Robotics* **2019**, *4*, eaau5872.
13. Jahanian, A.; Chai, L.; Isola, P. On the "steerability" of generative adversarial networks. International Conference on Learning Representations, 2020.
14. Shen, Y.; Gu, J.; Tang, X.; Zhou, B. Interpreting the latent space of GANs for semantic face editing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9243–9252.
15. Locatello, F.; Poole, B.; Rätsch, G.; Schölkopf, B.; Bachem, O.; Tschannen, M. Weakly-supervised disentanglement without compromises. *Journal of Machine Learning Research* **2021**, *22*, 1–66.
16. Love, E.; Tennakoon, B.; Maroulas, V.; Carlsson, G. Topological Convolutional Layers for Deep Learning. *Journal of Machine Learning Research* **2023**, *24*, 1–35. Submitted 1/21; Revised 2/23; Published 2/23.
17. Shi, Y.; Li, G.; Qin, Q.; Zhang, K.; Lin, Y.; Xiang, Y.; Ding, Y.; Lin, L. Contrastive learning for unpaired image-to-image translation. European Conference on Computer Vision. Springer, 2020, pp. 319–335.
18. Gaier, A.; Ha, D. Weight agnostic neural networks. Advances in Neural Information Processing Systems, 2019, Vol. 32.
19. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **2019**, *1*, 206–215.
20. Bostrom, N.; Dafoe, A.; Flynn, C. Public Policy and Superintelligent AI: A Vector Field Approach. In *Ethics of Artificial Intelligence*; Liao, S.M., Ed.; Oxford University Press, 2020.
21. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57.
22. Rahwan, I.; Cebrian, M.; Obradovich, N.; Bongard, J.; Bonnefon, J.F.; Breazeal, C.; Crandall, J.W.; Christakis, N.A.; Couzin, I.D.; Jackson, M.O.; others. Machine behaviour. *Nature* **2019**, *568*, 477–486.
23. Shih, A.; Sawhney, A.; Kondic, J.; Ermon, S.; Sadigh, D. On the Critical Role of Conventions in Adaptive Human-AI Collaboration. International Conference on Learning Representations (ICLR). ICLR, 2021.
24. Hospedales, T.; Antoniou, A.; Micaelli, P.; Storkey, A. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *44*, 5149–5169.