[1]Arjun Sirangi

Ethical Guardrails for Real-Time Generative Targeting Guardrails

Journal of Electrical Systems

Abstract: - Real-time generative targeting, powered by AI and machine learning algorithms, has revolutionised digital content, ad, and user experience personalisation. Real-time analysis of user activity, preferences, and contextual data allows these systems to dynamically provide tailored messages. Even while this technology increases participation and efficiency, we must address its major ethical issues. Algorithmic discrimination, user privacy violations, and manipulative content that exploits psychological flaws worry many individuals. Lack of solid regulatory frameworks and ethical monitoring increases the risk of these technologies being utilised for profit or politics. Thus, we must establish ethical boundaries for innovation and responsibility.

Keywords: Generative AI, Real-Time Targeting, Ethical Guardrails, Data Privacy

1 INTRODUCTION

There is tremendous ethical danger and revolutionary promise in the era of sophisticated artificial intelligence's real-time generative targeting, wherein AI systems generate user-specific content, ads, or communications in real-time based on their actions and data. Privacy, manipulation, algorithmic bias, and transparency are major issues that arise from this kind of targeting, even if it may maximise user engagement and business success

1.1. Problem Statement

Real-time generative targeting technologies driven by AI have progressed faster than ethics and regulations. These technologies may rapidly customise information based on user activity and data patterns without user consent. This raises the risk of user privacy breaches, computational biases, and psychological manipulation, especially among vulnerable populations. Accountability and ethical audits are difficult with opaque AI decision-making systems.

1.2. Research Objectives

1. For the purpose of safeguarding user information in AI systems that are constantly learning and adapting.
2. To ensure that generative targeting procedures are open and that participants are well informed.
3. Against the creation of biassed or manipulative material that takes advantage of user weaknesses.
4. To encourage transparency and equity by means of legal and technical protections.

2. BACKGROUND AND FOUNDATIONAL CONCEPTS

2.1. Privacy and Consent in Real-Time Data Collection

Behavioural, biometric, and contextual data streams are frequently used for real-time generative targeting. In the context of rapidly developing AI systems, this section delves into the question of how to safeguard individuals' rights to control their own data, reduce invasive surveillance, and guarantee informed consent.

2.2. Preventing Manipulation and Behavioral Exploitation

When generative systems tailor content for maximum impact, they can inadvertently cross ethical boundaries—nudging users toward decisions they might not otherwise make. This section examines how to recognize and limit psychological manipulation, particularly in vulnerable populations such as children or individuals with cognitive impairments.

[1] Advance Analytics Manager

2.3. Algorithmic Fairness and Accountability

It all comes down to the quality of the data and reasoning used to train generative AI systems. Implementing auditing, interpretability, and human supervision procedures to assure accountability across varied user groups is the emphasis of this subsection, which aims to eliminate bias in real-time targeting systems.

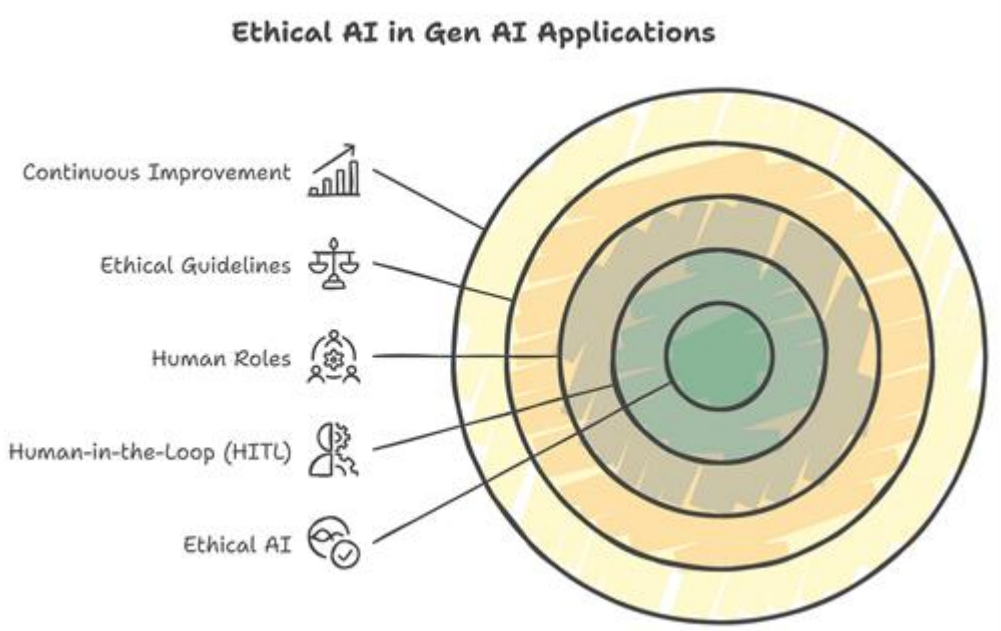


Fig . 1 Ethical AI in Gen AI Applications

3. REAL-TIME GENERATIVE AI APPLICATIONS

Among its several applications is real-time generative AI, which generates material in real-time. Demonstrating its adaptability, it drives customer service chatbots and aids in the creation of innovative content. To maximise the use of generative AI applications in real-time, we must understand its capabilities and limitations. Having a well-rounded perspective allows us to utilise it in innovative and thrilling ways.

The fundamental concepts, benefits, and challenges of real-time generative AI, as well as its applications across several sectors, will be examined in this blog article.

3.1 Challenges of Real-Time Generative AI

Latency and Response Time

- Real-time applications necessitate quick reaction times. When using sophisticated maths, a Generative AI application that generates content could slow things down and make real-time use problematic.
- How to make things go faster: You may speed up answers by making models smaller, removing unneeded pieces, and employing specific hardware.
- In order to cut inference time by 40-60%, researchers optimised a large-scale generative AI model for TPUs.

Computational Resources

Resource-hungry models: Models that need a lot of resources: Generative AI apps that create fresh, substantial material require a lot of processing power for learning and operation.

More gear: The current state of computer processing units (CPUs, GPUs, and TPUs) poses a challenge to the scalability and complexity of AI applications implemented in real time.

Using the cloud: When additional processing power is required, users can tap into cloud systems. It may take hundreds of GPUs to train a large-scale generative AI model, according to a study by OpenAI.

Data Limitations

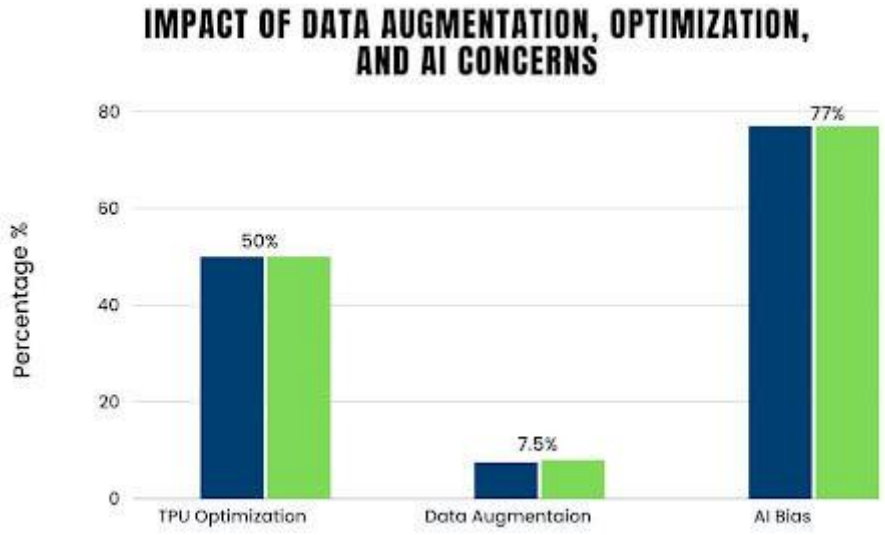
- Data quality and quantity: The performance of generative artificial intelligence models is highly impacted by both the quality and quantity of the training data.
- Data privacy: People may experience concerns regarding their privacy when large datasets are gathered and used.
- Data augmentation: The use of techniques such as augmentation can assist in overcoming data limitations and improving the performance of models in a variety of contexts.

3.2 Ethical Considerations

Bias and fairness: It is possible for generative artificial intelligence models to continue to transmit biases from their training data, which can result in outputs that are unjust or biassed.

Misinformation and deepfakes: The fact that generative artificial intelligence systems are capable of producing fake material that appears to be extremely real causes people to be concerned about deep fakes and incorrect information.

Transparency and explainability: For the purpose of ensuring that these systems are accountable and correcting any possible biases, it is essential to have a solid understanding of how generative AI models generate decisions. 77% of respondents to a study conducted by the Pew Research Centre expressed worry over the possibility of prejudice in artificial intelligence systems.



3.3 Techniques for Optimizing Real-Time Performance

Model Optimization

Pruning is the process of reducing the size of a model and making it easier to calculate by eliminating extraneous links and weights from the model. This is done in order to lower the size of the model.

Quantization: Reducing the precision of the numerical representations in the model in order to save both space and time for the necessary computations.

Distillation: The transformation of information from a large and complex model to a model that is more compact and effective.

- Convolutional neural networks may be pruned to shrink their size by as much as 90 percent without suffering a substantial loss in accuracy, according to a study conducted by Google AI.

- Quantisation has the potential to reduce the size of a model by as much as 75% while having a respectable level of accuracy.

Hardware Acceleration

GPUs: In deep learning, matrix operations and other calculations are frequently performed using graphics processing units (GPUs), which are computers that work in parallel to speed up these processes.

TPUs: Tensor Processing Units are pieces of hardware that are specifically designed for machine learning operations and give significant performance enhancements for particular jobs.

Cloud-Based Infrastructure

Scalability: With cloud-based platforms, resources may be scaled up quickly to meet the requirements of real-time applications.

Cost-efficiency: Costs may be optimised for fluctuating workloads with the aid of pay-as-you-go pricing.

Managed services: Service providers in the cloud offer services to handle artificial intelligence and machine learning, which makes it simpler to implement and administer.

- 80 percent of companies, according to a survey conducted by McKinsey & Company, employ cloud-based platforms for the development of artificial intelligence.

- AI systems that are hosted in the cloud have the potential to save development time by 30–40% and enhance time-to-market.

Efficient Data Pipelines

- This method processes data in batches in order to obtain a better throughput. Batch processing is a methodology.

- Streaming processing is a method that processes data in real time as it is received, and the phrase "streaming processing" refers to this method.

4. REAL-WORLD APPLICATIONS

4.1 Generative AI applications have an impact on many industries. Here are some standouts:

- **Healthcare:**

- A novel drug candidate is developed through the process of drug discovery, which involves the use of desirable features.

- Medical image analysis, which entails the production of fake medical images for the goal of training artificial intelligence models and expanding datasets by include more data.

- Applications of generative artificial intelligence have an impact on the process of drug development, making it thirty percent more productive, according to a study that was published in Nature Communications.

- **Entertainment:**

- The development of lifelike characters, settings, and narratives for use in the production of video games.

- Composing music, which involves producing music that sounds original and may be written in a number of styles.

- **Marketing and Advertising:**

- The process of designing products involves coming up with new ideas for objects and determining how they should seem to the consumer.

- A research that was carried out by McKinsey & Company found that the applications of generative artificial intelligence have the potential to increase the effectiveness of marketing efforts by a factor of two or even three.

4.2 Success Stories and Challenges Faced

- Success Story: This powerful text-to-image model, which is known as OpenAI's DALL-E 2, is able to produce pictures that are both lifelike and innovative. This demonstrates how applications of generative artificial intelligence have the potential to change the art and design sector.

- Challenge: It is essential to have high-quality, diverse training data in order to ensure that generative artificial intelligence application models function well.

- Success Story: An application for the development of landscapes, GauGAN was created by NVIDIA. Urban planners and architects make use of it in order to generate realistic pictures of projects that are currently in the planning phases.

Industry-Specific Applications

- E-commerce: Creating personalised marketing campaigns and writing product descriptions that involve offering recommendations for other items are two examples of similar activities.

- Finance: Producing synthetic financial data is being done in order to train fraud detection algorithms and evaluate risk. This is being done in order to evaluate risk.

- Education: In the process of developing personalised educational resources and evaluation standards.

- Manufacturing: improving the design of the product and making the production processes easier to learn.

Using the potential of generative artificial intelligence, organisations in a wide range of sectors are able to discover new ways to grow their operations, boost their efficiency, and make their customers more happy. Other benefits include increased customer satisfaction.

4.3 Best practices for establishing effective guardrails for generative AI

For the purpose of ensuring that your content consistently fulfils quality, brand, and compliance standards, it is vital to build guardrails for generative artificial intelligence. This requires rigorous planning and the use of the relevant technologies. Acrolinx should be used in accordance with the following best practices in order to ensure the establishment and maintenance of guardrails that are effective:

- **Create style guides**

When it comes to effective content governance, having a style guide that has been created for a long time is one of the most significant components. By utilising Acrolinx, you will have the ability to computerise your style standards and transform them into rules that can be used to drive the process of content production in real time.

For instance, if your organisation places a significant focus on having communication that is both clear and brief, Acrolinx will enforce sentence length constraints and give alternatives that are more intelligible for language that is verbose.

- **Regular compliance checks**

Maintaining compliance is not a one-time activity; rather, it calls for constant monitoring. Acrolinx automatically performs compliance checks by analysing material to determine whether or not it complies with regulations that are relevant to the industry.

For instance, Acrolinx may be utilised by financial institutions to ensure that product descriptions are in accordance with Consumer Duty legislation. Similarly, healthcare organisations can use it to verify that patient documents are in compliance with the standards of the FDA or HIPAA.

- **Create brand-specific terminology**

Consistent language helps to strengthen brand identification and promotes clarity, particularly in industries that are subject to a great deal of regulation. The creation of a centralised library of authorised brand-specific phrases is one of the ways that Acrolinx contributes.

Take for example a corporation that manufactures medical equipment and decides to standardise phrases such as "patient monitoring system" rather than using disparate names such as "monitor" or "PMS." Any variations from the authorised terms are brought to light by Acrolinx, which helps to reduce the likelihood of confusion or inadequate communication.

5. ETHICALLY-COMPLIANT GUARDRAIL DESIGN

54.1 Policies and rules

The establishment and maintenance of not just legal but also ethical standards that an organisation or a person desires to uphold depends on current conversational AI systems' adherence to rules and regulations. There has to be strong yet easy-to-understand systems in place to regulate content and conduct in light of the increasing number of AI assistants and conversational AI applications across many fields.

In order to provide organised but readily changeable enforcement inside customisable AI assistants, the suggested architecture classifies different rule kinds as policies. Two crucial ethical considerations are at play here. One option is for the organisation or service provider to pre-build a set of rules; another is for the user to add or alter additional rules to create a completely customised ethical guardrail.

Organisations concerned about AI safety or individual users might be the end users of these helpers. The employment of multiple large language models raises legitimate ethical concerns about privacy and security since LLMs learn and use the data they are fed as well. The goal of these regulations is to ensure that neither the LLM providers nor the end user are able to access any private or undesired information.

5.2 Types of rules

Three main types of rules exist, each with its unique technical complexity and strength. Users can utilise any or all of these capabilities to customise their rules to their ethical and privacy preferences.

AI assistants can discover and erase PII including email addresses, social security numbers, and phone numbers using static criteria and established patterns. Using a regex or similar pattern recognition tool in natural language processing (NLP) to detect or conceal personally identifying information may help prevent providing private information to third-party LLM providers.

Human-readable natural-language rules tell the user what to encourage or discourage in LLM conversations. Example: "avoid discussion about religion" or "never mention any content improper for children under 12". This rule type can handle several difficulties, such as industry-specific standards or keeping interactions courteous and free of offensive language. You may mix and match them with system prompts without affecting the original, making them unique. While system prompts only apply to AI output, natural-language standards may also be used for human input.

Two approaches to avoid using an LLM to decrease LLM responses are given below. Natural language processing rule enforcement is an advantage of NLP over LLMs. Lexicon-based methods, sentiment analysis, and user-defined descriptions coupled with specified keywords and phrases to avoid them are examples. Alternatively, a firm or sophisticated user might host their own open-source versions of popular language models like Llama or Mixtral to own the data.

5.3 Policies

Policies are composed of a combination of different regulations. Typically, policies' rules are checked for compliance in a sequential fashion. In order to immediately identify and filter out sensitive information based on clearly foreseeable patterns, static rules are applied initially. The following step is the enforcement of natural-language rules, which shape the discourse in a way that complies with the user-defined standards of

conduct. Finally, inputs and outputs are both categorised by trained classifier rules using learnt categories, enabling a more sophisticated comprehension and producing responses.

Remember that the user's privacy and ethical preferences determine the many ways a complying system may be customised; they can adjust this default sequence to construct a hierarchical chain of rules.

The AI output and the user's input can both make use of any of the policies.

5.4 Assistants

A combination of input/output rules, desired system prompts, and action items is used to construct an AI assistant. An input policy regulates the actions and enquiries that users are permitted to make in order to guarantee that they follow the principles and regulations of the organisation. An AI guardrail's principal function is to regulate the AI assistant's responses and interactions with humans in a way that satisfies user or business mandates.

Content that breaches a policy rule is considered inappropriate, and the system can carry out the user-configured action in such a case. The user is given the choice of what to do when a policy violation, defined as a breach of one or more rules, is detected. No sensitive data is ever communicated to an outside LLM provider, yet the user and AI system can continue communicating normally after redaction. Data disclosures of sensitive information, such social security numbers or trade secrets, can be erased if a rule detects them. The second potential issue is a blocking violation, which would put an end to the entire communication. Extra measures may be taken by companies, such as signalling a warning, recording the violation, or notifying a human. A warning message is sent to the user, informing them of the policy infraction and providing personalised remarks or advice based on their situation.

5.5 Ethical pluralism in AI design

Various ethical perspectives The necessity for adaptable barriers that can accommodate multiple ethical frameworks is highlighted by the fact that distinct persons and organisations may possess differing ethical viewpoints and principles. A fully aligned AI system may adhere to both an abstract moral model and varied standards and norms by letting people simply specify their values integrated into policies and regulations depending on their individual ethical concerns.

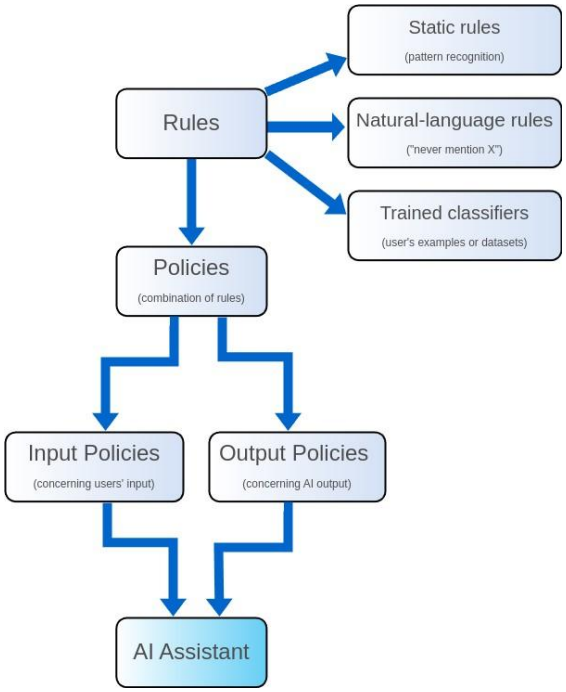


Figure 1: AI ethical management and enforcement hierarchy. Static rules, learnt classifiers, and natural language rules create policies in this image. To ensure ethical and operational compliance, AI assistants have these policies.

The design of AI systems should prioritise human aims and be consistent with human ideals, rather than blindly following predetermined goals that may be at odds with human principles. Here are Russell's three guiding principles:

- The sole purpose of the machine is to achieve to the greatest extent possible the fulfilment of human desires.
- Initially, the machine does not have a clear understanding of whatever those preferences are.
- The behaviour of individuals is the most reliable source of knowledge on human preferences.

The guardrails method in this work places an emphasis on practical mechanisms for establishing ethical norms in AI systems, whereas this approach is centred on the underlying principles of artificial intelligence ethics. The framework places an emphasis on rules and regulations that can be customised, so enabling users to set ethical norms that are relevant to their own beliefs and situations. This acknowledges the various ethical viewpoints that are not well addressed by Russell's approach, which is quite imprecise.

6. POTENTIAL DISAGREEMENTS

There is a possibility that there will be conflicts between the many guardrails that are contained inside a framework that has been outlined. When addressing the intricacies of guardrail conflicts within AI ethics systems, it is essential to recognise the many circumstances in which such conflicts may develop and the means by which they may be controlled. This is because different guardrail conflicts can arise in different ways. In this part of the article, we will investigate a number of different instances of guardrail opposition, classify them according to the type of the conflict they involve, and suggest a number of different solutions for resolving these conflicts.

Case 1: Complete and Permanent Opposition

Under this situation, Guardrail A and Guardrail B are constantly at odds with one other, indicating that their ethical vectors, which reflect policy combinations, are diametrically opposed, with a dot product of -1. When two ethical imperatives are inherently incompatible, a scenario like this one arises. Some barriers may place an emphasis on complete transparency, while others may place an emphasis on complete privacy. Statistic analysis might reveal this kind of situation by highlighting the underlying resistance prior to deployment by examination of the mathematical correlations or logical requirements established by these rules.

Variant I: When Guardrails A and B are the sole active guardrails, the system becomes "ethically blind" since their mutual negation removes all ethical guidance. Because it completely removes the AI's ethical guiding, this circumstance is very troublesome.

Variant II: Depending on other guardrails, the system may continue to operate ethically as intended even when A and B aren't operational. Inconsistencies in ethical thinking may arise, nonetheless, due to the ongoing contradiction between A and B.

Variant III: Even though there are several guardrails, the system will be left without moral direction if all of them are engaged in mutual negation. This utter rejection sets the stage for an extremely unpleasant situation in which the AI acts without regard for ethical considerations.

Case 2: Permanent but Limited Disagreement

The dot product of Guardrails A and B is near to -1, such as -0.9, and they are typically opposite to each other, but not entirely so. The setting is reminiscent of the boring but tolerable political debates that occur, for example, between the two main parties. There is resistance, but weighted averaging does help get people to agree on anything. These instances might be found via static analysis, allowing for changes to reach a balanced ethical position.

Case 3: Conditional Opposition

Depending on the particular input or context, Guardrails A and B are only ever totally opposed to each other, meaning their dot product is occasionally negative 1. In certain cases, the ethical vectors may coincide, but in others, they may be at odds. As in Case 1, Variants I, III, and III are also applicable here.

Case 4: Conditional but Limited Disagreement

In this scenario, Guardrails A and B can have slightly different values, with a dot product of around -0.9. This is similar to a short-lived political dispute, in which there is resistance but it does not prevent people from coming together to find a solution. By utilising weighted averaging, such conflicts may often be handled within the current framework of the system.

7. CONFLICT RESOLUTION STRATEGIES

To manage these conflicts, we consider the following strategies:

Weighted Averaging System: Weighted average takes into account the strengths of several guardrails, allowing for sophisticated ethical reasoning in most situations, particularly in Cases 2 and 4. But in Cases 1 and 3, when total resistance could cause ethical paralysis, this method fails.

Strict Order or Hierarchy of Precedence: To overcome the drawbacks of weighted averaging, a tight hierarchy of priorities can be set up, so that higher-priority barriers take precedence when there is a conflict. Even though this stops people from being completely uneducated about ethical issues, it might lead to unequal power dynamics if weakly-held views from high-precedence guardrails get a foothold.

Hybrid Approach: Presence Subject to Conditions. Weighted averaging might be the default technique in a hybrid approach, but strict order of precedence would be reverted to in cases when mutual negation is found, such as in Case 1/I, 1/III, 3/I, or 3/I. This system would be able to handle short-term disputes well while making sure that long-term oppositions cause alarms, informing users that the system is following limited ethical guidelines.

Contextual Triggering. The AI system may be programmed to implement various safety measures according to the situation at hand. In cases where sensitive personal information is involved, for instance, the privacy guardrail might be engaged, superseding the transparency requirement. On the other hand, the transparency guardrail might be more important in cases when public accountability is required. This method enables the instantaneous settlement of disputes by analysing the current state of affairs.

User resolution. When resolving conflicts automatically becomes difficult or when both guardrails are equally important, the system has the capability to alert humans to the disagreement. After then, administrators or users can choose by themselves which guardrail is more important in that particular scenario. Using the person-in-the-loop concept, this method is great for high-stakes situations that require human judgement with subtlety.

In order to create reliable AI systems, it is crucial to comprehend and resolve guardrail conflicts. Although Cases 2 and 4 include common and easily handled ethical disputes, Cases 1 and 3 provide far more substantial obstacles that need careful planning and methods for resolving conflicts. We can design systems that can handle complicated, competing directions while yet being ethical by combining weighted averaging with conditional precedence. With this method, we can be certain that AI systems will continue to act ethically and in accordance with a wide range of human values.

8. FUTURE TRENDS AND CHALLENGES

8.1 Emerging Technologies and Techniques

- Hybrid models: The combination of generative artificial intelligence applications with other methodologies, such as neural-symbolic AI and reinforcement learning, in order to construct models that are more robust and adaptable.

- Multimodal generative AI applications: The development of models that are capable of producing material in a variety of formats, including text, images, and sound.

- Explainable AI: Increasing the level of openness and knowledge of models that use generative artificial intelligence in order to gain confidence and solve ethical problems.

It is anticipated that hybrid artificial intelligence models would account for fifty percent of all applications of artificial intelligence by the year 2022, as stated by a survey conducted by McKinsey & Company.

8.2 Ethical and Responsible Development

- Bias reduction: For the purpose of ensuring that everyone is treated fairly and equally, we are working to eliminate biases in datasets and AI models.
- False information and synthetic media: Methods are now being developed for the purpose of recognising harmful material and putting a stop to the creation and spread of similar content.
- Data protection and system safety: with the goal of preventing unauthorised access to artificial intelligence platforms and securing sensitive information regarding those platforms.

Protecting sensitive information about artificial intelligence systems and blocking unauthorised access to those platforms simultaneously.

9. CONCLUSION

A number of sectors and parts of society stand to benefit greatly from the fast-developing area of generative AI applications. Generative AI is finding more and more uses in fields as varied as drug discovery, natural language processing, and the creation of photorealistic images and movies. The advantages of generative AI applications much outweigh the difficulties, which include issues of ethics and computing power.

Using technology to our advantage may propel innovation, boost efficiency, and solve critical social problems. We may anticipate many more revolutionary uses of generative AI in the years to come as research and development keep pushing the field forward. Responsible use of this technology is of the utmost importance, as is the assurance that its advancement is in line with moral standards and community ideals.

REFERENCES

[1] Ackerman, M. S., Dachtera, J., Pipek, V., & Wulf, V. (2013). Sharing knowledge and expertise: The CSCW view of knowledge management. Computer Supported Cooperative Work (CSCW), 22(4–6), 531–573. https://doi.org/10.1007/s10606-013-9192-8

[2] Alzheimer's Research UK. (2022). Explaining the amyloid research study controversy. https://www.alzheimers.org.uk/for-researchers/explainingamyloid-research-study-controversy

[3] Bao, C., Li, S., Flores, S., Correll, M., & Battle, L. (2022). Recommendations for visualization recommendations: Exploring preferences and priorities in public health. arXiv. https://arxiv.org/abs/2202.01335

[4] Knorr Cetina, K. (1999). Epistemic cultures: How the sciences make knowledge. Harvard University Press.

[5] Chan, J., Ding, Z., Kamrah, E., & Fuge, M. (2020). Formulating or fixating: Effects of examples on problem solving vary as a function of example presentation interface design. arXiv. https://doi.org/10.48550/arXiv.2401.11022

[6] Chen, Z., & Xia, H. (2022). CrossData: Leveraging text-data connections for authoring data documents. In CHI Conference on Human Factors in Computing Systems (pp. 1–15). ACM. https://doi.org/10.1145/3491102.3517485

[7] Chen, Z., Xiong, Z., Yao, X., & Glassman, E. (2020). Sketch then generate: Providing incremental user feedback and guiding LLM code generation through language-oriented code sketches. arXiv. https://arxiv.org/abs/2405.03998

[8] Ding, Z. (2020). Advancing GUI for generative AI: Charting the design space of human-AI interactions through task creativity and complexity. In Companion Proceedings of the 29th International Conference on Intelligent User Interfaces (pp. 140–143). ACM. https://doi.org/10.1145/3640544.3645241

[9] Ding, Z. (2020). Towards intent-based user interfaces: Charting the design space of intent-AI interactions across task types. arXiv. https://arxiv.org/abs/2404.18196

[10] Ding, Z., & Chan, J. (2020). Intelligent Canvas: Enabling design-like exploratory visual data analysis with generative AI through rapid prototyping, iteration and curation. arXiv. https://doi.org/10.48550/arXiv.2402.08812

[11] Ding, Z., Smith-Renner, A., Zhang, W., Tetreault, J. R., & Jaimes, A. (2019). Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation. arXiv. https://arxiv.org/abs/2310.10706

[12] Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019). Increasing the transparency of research papers with explorable multiverse analyses. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–15). ACM. https://doi.org/10.1145/3290605.3300295

[13] Y. Dong, R. Mu, G. Jin, Y. Qi, J. Hu, X. Zhao, J. Meng, W. Ruan, and X. Huang, "Building Guardrails for Large Language Models," arXiv preprint arXiv:2306.07500, 2010.

[14] Y. Wang and L. Singh, "Adding Guardrails to Advanced Chatbots," arXiv preprint arXiv:2306.07500, 2010.

[15] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, "Instruction Tuning for Large Language Models: A Survey," arXiv preprint arXiv:2308.10792, 2022.

[16] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning: A Survey," arXiv preprint cs/9605103, 2022.

[17] Y. Huang and Q. Zhu, "Deceptive Reinforcement Learning Under Adversarial Manipulations on Cost Signals," in International Conference on Decision and Game Theory for Security, pp. 217–237, Springer, 2019.

[18] V. Behzadan and A. Munir, "Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks," in International Conference on Machine Learning and Data Mining in Pattern Recognition, pp. 262– 275, Springer, 2017.

[19] T. Traian Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen, "NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails," arXiv preprint arXiv:2310.10501, 2022.

[20] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 2022.

[21] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, "Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment," arXiv preprint arXiv:2312.12148, 2021.

[22] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, and T. D. et al., "Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations," arXiv preprint arXiv:2312.06674, 2022.

[23] E. Mason, "Value Pluralism," in The Stanford Encyclopedia of Philosophy (E. N. Zalta and U. Nodelman, eds.), Metaphysics Research Lab, Stanford University, summer ed., 2022.

[24] M. Zheng, J. Pei, and D. Jurgens, "Is "A Helpful Assistant" the Best Role for Large Language Models? A Systematic Evaluation of Social Roles in System Prompts," arXiv preprint arXiv:2311.10054, 2021.