Q1. List the three types of learning algorithms and give a real-life example for each type.

There are three types of learning algorithms: supervised learning, unsupervised learning, and reinforcement learning.

A supervised learning algorithm learns a function from provided examples of its inputs and corresponding outputs. Specifically, the algorithm is provided with labelled training data, which it used to determine the correct output. In fact, e-commerce companies such as Amazon, and streaming platforms such as Netflix employ supervised learning algorithms to customise product recommendations that enhance sales, customer engagement, and loyalty. For instance, Amazon.com uses historical purchase and browsing data as inputs, which are labelled with whether or not the user purchased the item as outputs. Based on this labelled dataset, machine learning algorithms are trained to predict and recommend future purchases. As a result, 35 percent of Amazon.com sales and 75 percent of Netflix.com viewings are influenced by personalised product recommendations.

An unsupervised learning algorithm learns patterns from input data to uncover concealed structures or relationships within the data, without the specification of output values. In the real world, financial services companies such as Visa, Mastercard, and JPMorgan Chase practically employ unsupervised learning algorithms to detect fraudulent transactions on credit cards, which is also called Credit Card Fraud Detection. Specifically, unsupervised learning algorithms categorise transactions into either the 'normal' or 'suspicious' cluster based on factors like transaction amount, location, and timing. The classification is achieved without requiring explicit labelling to detect abnormal patterns or outliers that may indicate fraud in real-time. As a result, companies can reduce financial losses, increase trust, and protect their customers' accounts, since American fraud victims now number over $150 million, up from $127 million in 2021 [2].

A reinforcement learning (RL) algorithm makes decisions based on continuous past experiences and patterns, interacts with the environment by performing certain actions, and maximises cumulative rewards through trial and error. As Tesla utilises reinforcement learning to power its Autopilot feature, the self-driving technology can continuously reduce accidents, improve the driving experience, and reduce energy consumption. By leveraging the same data set collected from millions of miles driven by Tesla cars across the globe and real-time traffic data, Tesla’s Autopilot feature improves route planning, improves lane follow capabilities, and even automates parking. In this way, Tesla's autonomous driving system continuously improves its driving performance based on real-world data using reinforcement learning. Over time, RL algorithms receive feedback in the form of rewards or penalties for their actions, enabling them to learn and make better decisions.

Reference:

- [1] MacKenzie, I., Meyer, C., & Noble, S. (2013, October 1). How retailers can keep up with consumers. McKinsey & Company. https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers
- [2] Credit Card Fraud 2021 Annual Report: Prevalence, Awareness, and Prevention. (n.d.). Security.org. https://www.security.org/digital-safety/credit-card-fraud-report/

Q2. List 2 most important ethical issues in AI in your own view and give your reason why.

Artificial Intelligence (AI) has progressively integrated into a major part of modern society. It has shown remarkable capabilities in a wide range of fields in the modern world. However, alongside rapid advancements in artificial intelligence, many raised critical ethical concerns that require immediate attention from these systems. The two most important ethical issues in AI, in my opinion, are **Autonomy & Accountability**, and **Data Privacy and Security**.

Firstly, a crucial ethical issue is **Autonomy & Accountability**. Any AI system has the ability to carry out decision-making independently without human intervention which is known as Autonomy. Therefore, AI has great potential to optimise processes and reduce human error. However, concerns are raised, particularly on fairness, transparency, and **accountability**, particularly when an autonomous AI system makes a decision that leads to harm or breaches ethical standards. An example is when autonomous vehicles cause fatal accidents, the question on responsibility arises. The ethical implications of autonomous systems are becoming increasingly apparent, especially in the healthcare industry, where decisions can have a life-or-death impact (van der Waa et al., 2021) [3].

When decisions are made by machines, the issue of accountability becomes more complex, challenging traditional notions of ethics and responsibility. Lack of clear accountability makes it difficult to enforce ethics and raises questions regarding whether existing laws and regulations are sufficient to hold someone (or something) accountable. As suggested in a study published in the BMC Medical Ethics journal (Durán & Jongsma, 2021) [4], using black box algorithms in medicine raises concerns about transparency and ethical issues, but their reliability makes them trustworthy for medical applications. Despite this, the authors emphasise that trustworthiness alone is not sufficient; ethical deliberation and interdisciplinary transparency collaboration are fundamental to the responsible use and implementation of these algorithms.

An additional ethical concern is Data Privacy and Security, since there is substantial reliance of AI systems on large datasets containing confidential personal information; for example, Electronic Health Records (EHRs) or Personally Identifiable Information (PII) in the financial services sector (FinTech).

An essential objective of data privacy and security of artificial intelligence is to mitigate unauthorised access, use, and disclosure of private and sensitive information to ensure ethical usage of any system. As data breaches becomes increasingly common, handling Data Privacy and Security ethically becomes more important than ever. This is because a misuse of data can lead to a loss of trust, a violation of privacy, or can have serious legal implications. The issue is significant as the potential for misuse of personal data becomes more recurrent. For example, the use of AI algorithms can be exploited to predict political affiliations, sexual orientation, or even medical conditions. These predictions can be exploited for targeted advertising, political campaigns, or even harmful purposes.

Several workaround solutions are available to mitigate these risks, such as robust encryption methods, data protection, and strict data usage policies. For example, an identity-based secure and encrypted data-sharing technique has been proposed for improving data security in cloud-based eHealth systems (Sivan & Zukarnain, 2021) [5]. Furthermore, by utilizing Federated Learning for privacy-preserving (Dash et al., 2022) [6], data can be decentralized, thereby reducing the risk of a data breach. This innovative approach addresses cybersecurity concerns and increases FinTech growth while also maintaining data privacy and security.

In summary, both autonomy and accountability, as well as data privacy and security, are critical ethical issues in artificial intelligence that require immediate attention from all stakeholders. A comprehensive multi-disciplinary approach that incorporates technological solutions, regulatory frameworks (such as European guidelines on ethics in artificial intelligence), law, and ethical auditing can offer a balanced approach. Therefore, it is vital that ethical considerations develop alongside the rapid evolution of artificial intelligence to ensure responsible usage of technology in the future.

Reference:

[3] van der Waa, J., Verdult, S., van den Bosch, K., van Diggelen, J., Haije, T., van der Stigchel, B., & Cocu, I. (2021). Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations. Frontiers in Robotics and AI, 8. https://doi.org/10.3389/frobt.2021.640647

[4] Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. Journal of Medical Ethics, 47(5), medethics-2020-106820. https://doi.org/10.1136/medethics-2020-106820

[5] Sivan, R., & Zukarnain, Z. A. (2021). Security and Privacy in Cloud-Based E-Health System. Symmetry, 13(5), 742. https://doi.org/10.3390/sym13050742

[6] Dash, B., Sharma, P., & Ali, A. (2022). Federated Learning for Privacy-Preserving: A Review of PII Data Analysis in Fintech. International Journal of Software Engineering & Applications, 13(4), 1–13. https://doi.org/10.5121/ijsea.2022.13401

Q3.1. What is the main motivation of AI Safety with respect to DNN? For the work of SVDNN, summarize in your own words their main contribution and one limitation.

A fundamental motivation for AI Safety within the context of Deep Neural Networks (DNNs) is to ensure the reliability, security, and ethical functioning of these complex models. This is because DNNs are being deployed progressively in applications that needs to ensure a high level of safety, such as in industries of healthcare, autonomous vehicles, and financial systems. Errors or malicious behaviour in these industries can lead to serious consequences. Therefore, while the power and decision-making process of DNNs are considerable; they often appear to be "black boxes". This means it's often difficult to understand their decision-making process. Therefore, concerns are raised about the lack of transparency and interpretability in DNNs for the likely vulnerability to adversarial attacks, data poisoning, and unintended biases. As a result, AI Safety strives to develop comprehensive methodologies to verify, validate, enhance robustness, and monitor continuously DNNs to ensure that they the desired safety and ethical standards are met, especially within high-risk environments.

The field of image classification are witnessing the revolution from Deep Neural Networks. Specifically, they are achieving impressive experimental results that are comparable to the cognitive abilities of humans when performing complex tasks. However, they may be surprisingly unstable with respect to adversarial perturbations, which are minor changes to the input image, such as scratches, changes in camera angle, or changes in lighting conditions, that cause the network to misclassify the image. This raises concerns about the safety of these networks and the correct behaviour of a machine learning component, particularly in applications such as perception modules and end-to-end controllers for self-driving cars.

The purpose of Safety Verification of Deep Neural Networks (SVDNN) [7] is to enhance the safety of deep neural networks by developing a novel automated verification framework for feed-forward multi-layer neural networks. The automated verification framework for verifying the safety of DNNs, particularly focusing on classification tasks, systematically explores regions around a data point to search for adversarial manipulations of a given type, and propagates the analysis into deeper layers. This approach was implemented using Satisfiability Modulo Theories (SMT) and validated on several state-of-the-art neural network classifiers for realistic images.

There is a major limitation to the system in terms of exponential complexity and scalability. The verification process is computationally intensive, exponential in the number of features, and prohibitively complex for larger images. Further, although parallelization can improve efficiency and scalability, the current method has prohibitive complexity for large-scale applications.

Q3.2. List another important work in AI Safety and why you think it is important. (Give the proper citation and a link to its publication.

Formal Specification for Deep Neural Networks [8] is another significant contribution to the field of AI Safety. It explores the use of formal specification methods for designing and verifying DNNs to ensure AI safety and reliability. A formal specification is essential for deep neural networks as it is a set of well-defined rules that outlines the expected behaviour of a system. The article classifies specifications that are useful for reasoning about neural networks and the systems that use them within the context of artificial intelligence. Further, it highlights the importance of specifications in verifying, testing, retraining, and capturing design assumptions. The various methods for measuring neural network robustness are classified into safety properties, hyperproperties, probability properties, and coverage properties. It is possible to reason in depth about the behaviour and properties of DNNs using formal specification techniques, including temporal logic, probabilistic programming, and reinforcement learning with temporal logic constraints.

A formal specification of deep neural networks is important for a number of reasons:

- Robustness: AI systems can often be highly susceptible to adversarial attacks, where even minor changes to the input can lead to incorrect results. A formal specification can be useful in identifying these vulnerabilities in advance.
- Ethical and Legal Compliance: As AI systems are increasingly used in critical applications such as healthcare and autonomous vehicles, formal specifications are essential for ensuring that they are in compliance with ethical and legal standards.
- Debugging and Verification: Formal specifications allow for a more rigorous verification process, which ensures that the system behaves as expected under all conditions, since traditional debugging methods are often insufficient for complex AI systems.
- Interdisciplinary Collaboration: The paper references work in political philosophy and probabilistic programming, which suggests that solving AI safety is not just a technical challenge but also requires insights from other disciplines.

Despite AI's significant impact on many aspects of human life, there are challenges and vulnerabilities associated with applying formal specification techniques to DNNs. These potential challenges include complexity, scalability, and interpretability. The paper attempts to address the challenges by suggesting a more systematic design methodology for DNNs, further highlighting how safety concerns should be addressed in AI systems; along with providing insights into how formal methods can be used to guarantee the reliability and safety of AI systems. This work may potentially have a major impact on the development of AI systems that are ensured to have more reliability, trustworthiness, and alignment with human values.

Q4.1. What is the main motivation of XAI? For the work of LIME, summarize in your own words their main contribution and one limitation.

As a solution to the lack of transparency and "black boxes" associated with complex machine learning models, US DARPA (The Defense Advanced Research Projects Agency) has launched Explainable AI (XAI). The program aims to produce more explainable models while maintaining high prediction accuracy and enabling human users to understand, trust, and manage AI systems effectively.

The XAI program aims to develop more explainable and interpretable models that are both accurate and interpretable, since Machine Learning (ML) models become increasingly complex, their decision-making processes become less transparent. Additionally, another goal is to facilitate Human-AI collaboration in order to understand, be trustworthy, and effectively manage AI systems. It is of vital importance in mission-critical applications such as healthcare, finance, and defence.

The article "Why Should I Trust You?" [9], which presented LIME (Local Interpretable Model-agnostic Explanations) is an important contribution to the field of XAI. LIME proposed a novel explanation technique for explaining predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction.

LIME aims to explain the individual predictions of any classifier in a model-agnostic way. In order to accomplish this goal, it approximates the complex model with a simpler, interpretable model that is locally faithful to the complex model's decisions. Users can thus gain a greater understanding of the AI system's reasoning behind a particular decision. For example, LIME may be able to provide explanations for why a patient was classified as high risk for a certain disease, identifying the medical diagnosis responsible for the classification. A clinician who needs to understand the reasoning behind the model in order to make an informed decision will find this level of detail highly valuable.

Despite LIME's excellent performance in explaining individual predictions, its local nature may not provide a comprehensive understanding of the overall behaviour of the model. There could be a problem in applications requiring a detailed understanding of global behaviour. Nevertheless, LIME represents a pioneering effort to make AI more understandable and trustworthy, aligning well with DARPA's XAI program.

Q4.2. List another important work in XAI and why you think it is important. (Give the proper citation and a link to its publication)

Explainable Artificial Intelligence (XAI) in deep learning-based medical image analysis [10] is an important contribution to XAI research. Transparency, accountability, and explain-ability are becoming increasingly important in AI/ML models, particularly in the fields of XAI and medical imaging. It is not only a legal requirement, but also an ethical obligation to ensure that AI technologies are deployed in healthcare responsibly.

The paper emphasises the importance of XAI for fulfilling the requirements of ethical and regulatory compliance, such as the European Union's General Data Protection Regulation (GDPR), which mandates that patients have the right to a detailed understanding of the rationale behind automated decision-making processes affecting them. A study published in the paper found that 77% of physicians were more likely to correctly diagnose chest X-rays with an XAI providing a visual explanation when compared to chest X-rays without an XAI, emphasising the real-world impact of XAI in healthcare. It is more than a statistic; it is a practical clinical tool that can save lives and optimise healthcare resources. Furthermore, three criteria are used to distinguish XAI techniques: model-based versus post-hoc, model-specific versus model-agnostic, and global versus local (i.e., the scope of the explanation). Future directions for XAI in medical image analysis may include biological explanations derived from imaging features. Medical imaging may be revolutionised, allowing researchers to gain new insights into disease mechanisms and even enable personalised medicine.

In conclusion, as artificial intelligence technology evolves and integrates deeper into healthcare, it is essential to provide a timely and valuable guide for the responsible and effective use of more interpretable and explainable AI.

References

- [7] Huang, X., Kwiatkowska, M., Wang, S., & Wu, M. (2017). Safety Verification of Deep Neural Networks. Computer Aided Verification, 3–29. https://doi.org/10.1007/978-3-319-63387-9_1
- [8] Seshia, S. A., Desai, A., Tommaso Dreossi, Fremont, D. J., Ghosh, S., Kim, E. S., Sumukh Shivakumar, Vazquez-Chanlatte, M., & Yue, X. (2018). Formal Specification for Deep Neural Networks. Lecture Notes in Computer Science, 20–34. https://doi.org/10.1007/978-3-030-01090-4_2

- [9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. https://doi.org/10.1145/2939672.2939778
- [10] van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis, 102470. https://doi.org/10.1016/j.media.2022.102470