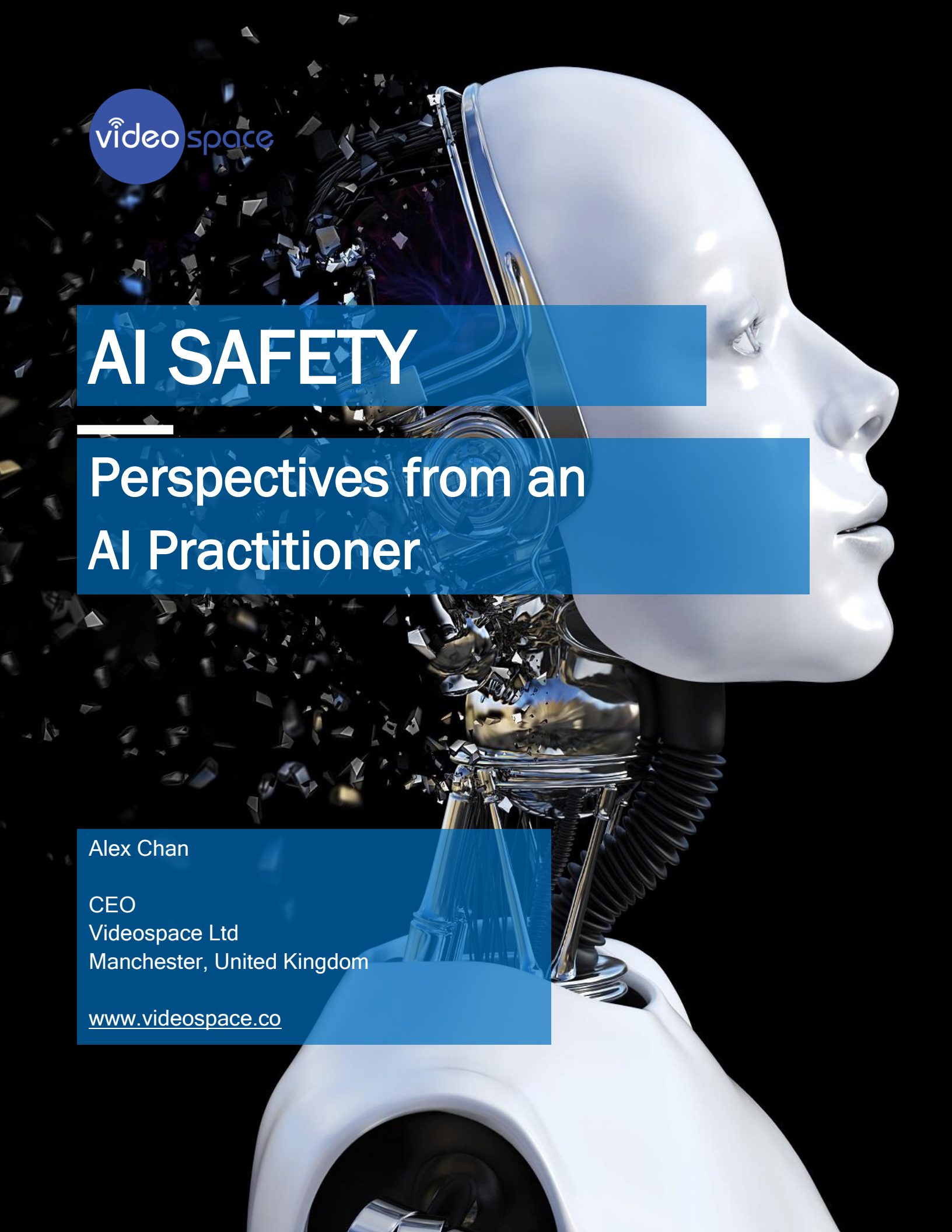

AI SAFETY

Perspectives from an AI Practitioner

Alex Chan

CEO
Videospace Ltd
Manchester, United Kingdom

www.videospace.co

Contents

Introduction

Artificial intelligence (AI) has the potential to transform our lives in many ways, but it also raises ethical concerns. As AI systems become more advanced and ubiquitous, it is increasingly important to ensure that they are used responsibly and ethically.

While lawmakers struggle to keep pace with global AI developments, sometimes, it is up to practitioners who are building, implementing and deploying AI systems to become gatekeepers.

Remember, AI has no conscience, people do.

There are many potential risks associated with AI, such as the risk of bias, the potential for AI to be used for malicious purposes, and the risk of unintended consequences. To mitigate these risks, it is important to develop and implement safety measures that can help ensure that AI systems are safe and reliable. Admittedly, sometimes these measures are the furthest thing from our minds.

Part 1 - AI Ethical Considerations

Some of the key areas of focus for AI safety include ethical considerations, risk assessment and mitigation, and best practices for designing, developing, and deploying AI systems. By taking a proactive approach to AI safety, practitioners can help ensure that AI is used in a way that benefits society as a whole.

Some of the key ethical considerations of AI safety includes:

1. **Bias**: AI systems can be biased if they are trained on biased data or if they are designed with implicit biases.

2. **Privacy**: AI systems can collect and process large amounts of personal data, which can raise privacy concerns. It is important to ensure that AI systems are designed with privacy in mind and that they comply with relevant data protection laws.

3. **Transparency**: AI systems can be difficult to understand and interpret, which can make it difficult to identify and correct errors or biases. It is important to ensure that AI systems are transparent and that their decision-making processes can be explained.

4. **Accountability**: AI systems can make decisions that have significant impacts on people's lives, but it can be difficult to assign responsibility for these decisions. It is important to ensure that AI systems are designed with accountability in mind and that there are mechanisms in place to address any negative impacts.

5. **Safety**: AI systems can pose risks to people's safety if they are not designed, developed, and deployed in a safe and responsible manner. It is important to ensure that AI systems are rigorously tested and that they comply with relevant safety standards.

These are just a few of the ethical considerations of AI safety. As an AI practitioner, it is important to be aware of these considerations and to take them into account when designing, developing, and deploying AI systems.

Part 2 - Risks associated with AI

As with all new tools and technologies, there are risks associated. Especially with AI as it brings along two characteristics: **scale** and **ubiquity**.

AI will be so common that it'll be invisible.

We will be relying on AI in our work and daily life, to increase our productivity, influence our decisions, and in many cases, even making our decisions for us. Therefore, it is important to know and understand some of the risks associated with AI:

1. **Discrimination**: AI systems can perpetuate or amplify societal biases due to biased training data or algorithms. This can lead to unfair treatment of certain groups of people. I would even go so far to say that it's impossible to eliminate discrimination biases because its AI creators (humans) are fundamentally biased. Thus, the question really is to understand and attempt to minimize this risk.

2. **Privacy**: AI systems can collect and process large amounts of personal data, which can raise privacy concerns. Data protection laws are different everywhere because attitudes are different everywhere. Moreover, different countries are at different stages of development. But it is important to ensure that AI systems are designed with privacy in mind and that they comply with relevant data protection laws.

3. **Safety**: AI systems can pose risks to people's safety if they are not designed, developed, and deployed in a safe and responsible manner. This is especially if the AI can directly influence our decision-making or physical safety. It is important to ensure that AI systems are rigorously tested and that they comply with relevant safety standards.

4. **Jobs**: AI-powered automation is a pressing concern as the technology is adopted in industries like marketing, manufacturing, and healthcare, etc. Job losses are going to be inevitable. At personal, societal and even global levels, it is important that we understand and prepare for these shifts, because they will be seismic. Thus, the question we should be asking ourselves is: How are we going to prepare for this shift?

As an AI practitioner, we know that AI can be blackbox of chocolates, where you never know what you're going to get. Therefore, it is all the more important to be aware of these risks and to take them into account when designing, developing, and deploying AI systems.

Part 3 - Best Practices for Ensuring AI Safety

While it's impossible to summarize best practices into a few short paragraphs covering all industries, I believe the following points can be used as starting points of a framework where we can start thinking about what we need to do as practitioners. Admittedly, some of these are easier said than done, and it also means more time, effort and resources. This is especially hard if the AI system has no precedence.

One of the difficulties that practitioners face is that customers or end-users often think AI is some sort of a silver bullet, one that delivers consistent results. While in truth, it doesn't always play out this way and AI safety can be the furthest thing from their minds.

1. **Design for Safety**: Safety should be a primary consideration when designing AI systems. This means identifying potential risks and hazards and designing systems that can detect and mitigate them. It also means designing systems that are transparent and explainable, so that their decision-making processes can be understood and audited.

2. **Test Rigorously**: AI systems should be rigorously tested to ensure that they are safe and reliable. This includes testing for potential biases, vulnerabilities, and other risks. It also means testing systems in a variety of scenarios and environments to ensure that they are robust and effective.

3. **Monitor Continuously**: AI systems should be monitored continuously to ensure that they are functioning as intended. This includes monitoring for potential errors, biases, and other risks. It also means monitoring systems for changes in their environment or inputs that could affect their performance.

4. **Collaborate Across Disciplines**: Ensuring AI safety requires collaboration across disciplines, including computer science, engineering, ethics, law, and policy. By working together, practitioners can identify potential risks and develop effective solutions to mitigate them.

5. **Stay Up-To-Date**: AI is a rapidly evolving field, and new risks and challenges are emerging all the time. Practitioners should stay up-to-date with the latest developments in AI safety and be prepared to adapt their practices as needed.

By following these practices and others like them, practitioners can help ensure that AI is used in a way that benefits society as a whole.

Part 4 - Future of AI Safety

There are a few areas that are really emerging along with the strong emergence of Generative AI. I believe the following areas are going to be very important because they will determine which direction or path AI will take globally.

1. **Global Regulation**: The UK AI Safety Summit, combined with a G7 declaration and US executive order, shows action is happening on AI safety. The Bletchley Declaration, a world-first global agreement at the UK's AI Safety Summit, was signed by 28 countries including the US, UK, and China, alongside the European Union, agreeing on the need for regulation. The US President Joe Biden has signed an executive order on artificial intelligence to outline how the country will regulate the technology. These are important landmarks, and there will be more to come. However, one challenge will be how regulations can be upheld, policed and executed when the rules are broken.

2. **Emerging Trends**: The field of AI safety has shed its status as the unloved cousin of the AI research world and took center-stage in 2023 for the first time. But amid a lack of global consensus on the way forward for regulation, developers of cutting-edge AI systems were "making a push to shape norms" by proposing their own regulatory models. Perhaps that's the way moving if governments and regulators are not able to come to agreement.

3. **AI Ethics**: As AI systems become more advanced and ubiquitous, it is increasingly important to ensure that they are used responsibly and ethically. Some of the key ethical considerations of AI safety include bias, privacy, transparency, accountability, and safety, are really coming to the fore as people start to realise the need for us to be aware of the good and bad with AI.

4. **AI Education**: Finally, what the world really need is more education about AI globally. This is to narrow the AI divide and level the playing field for less developed regions.

Conclusion

This concludes the four-part series on **“AI Safety: Perspective from an AI practitioner”**.

However, this is just the beginning and I am just seeing the world of AI from a practitioner's perspective. We do have to recognise that there are many perspectives. The nuclear threat is clear, we absolutely cannot allow AI to take over any form or command and control.

Mankind should not and cannot end by because of an algorithm.

That is why we need AI Safety!

Thus, it is important for various stakeholders to come together and start a conversation. Hopefully, this conversation will cut through geo-politics and various kinds of differences to make the world a safer place.

About Videospace

Mankind spent the last sixty years indexing and extracting invaluable data and intelligence from various digital formats (e.g. documents, webpages). However, one format still escapes us - videos. Video is the only format that Google cannot index and search. Therefore, many consider videos to be the last frontier of search technology.

Videospace's value proposition is our unique ability to **Search and Extract Video Data and Intelligence**. Using a multimodal AI approach, we created various world's first in Deep Video Search and developed one of world's most advanced Video Search Engine.

To find out more about Videospace, please go to www.videospace.co.