Cyberfort

WHITE PAPER

Taking the right steps to ensure 'Trust by Design' is embedded into your organisations AI strategy

Introduction

AI is a hot topic. Most organisations are discussing and reviewing AI as part of their business and technology strategies. It is estimated by the Office of National Statistics that 1 in 6 UK companies have adopted at least 1 AI tool into their businesses in 2023 (1). With Gartner discovering 55% of organisations are looking to adopt an AI first strategy moving forward (2) and projected spend on AI estimated to reach £16.8bn by the of 2024 in the UK according to the US Internation Trade Administration (3). It is clear from the research identified above AI adoption will continue to move at pace.

The business cases of AI have been well documented by many technology industry commentators over the past 12 months. For example, a significant amount has been written about the power of AI in terms of identifying how it can make their employees more productive, how organisations can unlock insights from the data they hold, improve abilities to complex solve problems quicker than ever before and respond to customers with more timely and accurate information.

But creating and designing the right AI strategy for an organisation is not an easy task. Many organisations are still in a reactive mode in terms of creating and implementing their AI strategy. Key considerations for an AI strategy to be effective need to include the skills available, security, infrastructure support, employee access, AI tool development and the design of a trustworthy AI model.

This article discusses why 'Trust by Design' needs to be included as part of an organisations AI strategy and how your organisation can take practical steps to make sure it is included moving forward.

What are the key areas organisations should focus on with 'Trust by Design' and AI

Trust in an organisation is crucial for success. It impacts relationships and outcomes with employees, investors, industry analysts, customers, suppliers and the technology used in an organisation. As an AI strategy may potentially touch on each of these stakeholder groups it is important Trust by Design is embedded in any AI strategy which is developed for an organisation.

Trust by Design has been in existence for many years and is seen as an approach for moving risk from a negative impact discussion topic to a more positive approach. By implementing a Trust by Design approach, a mindset shift can happen. It changes risk being viewed as just managing a set of factors which could impact the organisation financially, legally and reputationally to one which encourages innovation, reviewing where value can be unlocked, and how to engage in a more positive manner with customers, suppliers and key stakeholder groups.

So, what are the key areas your organisation should be focused on when embedding Trust By Design into an AI strategy?

Is AI being ethically used?: Any AI tool designed and deployed must comply with social and an organisations ethical standards. As part of the design stage of an AI tool, human behaviour needs to be factored into how the AI tool will be used and the purposes it will be used for. For example, will the data being used to inform LLM’s be used accurately and fairly, will users keep the data safe and compliant with industry regulations. All the data and the use of AI tools should be assessed in terms of moral behaviour, respect, fairness, bias and transparency.

AI usage needs to be socially responsible: The societal impact and potential of AI is well documented in terms of human wellbeing and financial effects it could have both positively and negatively. When designing and deploying AI as a key strategy organisations need to review the effects of AI on employees in terms of disrupting their working patterns, what skills will be required to make AI trusted and effective, the potential impacts on customers and other stakeholder groups if biases occur and the risks associated with these biases if AI is being used incorrectly.

Transparency, Explainability and Biases must be considered for trust to be truly embedded with AI: When dealing with Trust By Design for AI, transparency, explainability and bias are all crucial areas which must be considered, reviewed and appropriate policies implemented to keep users and the organisation safe and seen as trustworthy. End users need to understand if/when they are dealing with AI and be given options to choose how they interact with AI, how their data can be used as part of AI tools and the organisation needs to make sure right levels of data protection and consent are in place against industry and country guidelines for compliance purposes.

As well as being transparent organisations need to be able to explain how they are using AI in terms of the methodologies, training which has been put in place, data safeguards and decision-making criteria AI is using to produce results. Those whose data is being used for AI purposes will want to know it is processed in a fair and secure way and is not likely to be breached. To improve trust organisations should create a public set of guidance explaining how data is being kept, stored and securely managed for AI purposes. This in turn will give data subjects confidence that their data is being correctly used and not mismanaged or open to cyber-attacks.

Finally, and probably the issue which causes untrustworthiness is bias. AI and the results it produces is only as good as the data inputted. Inherent biases in data sets being used for LLM's informing AI systems could cause significant damage to a company's reputation if not identified and acted upon. For example, data biases could occur if the development team is only 1 or 2 individuals reviewing data or an LLM is sourced from an unverified 3rd party and has not considered a wide range of different viewpoints in terms of how data is being identified and used. Additionally, users of LLM's and AI systems may not be trained on how to spot specific biases in data and using data incorrectly in their AI decision making. When an AI strategy is being designed biases in data must be considered and a roadmap put in place to correct any biases which may be occurring due to AI.

Making AI accountable for the results it produces: As AI has exponentially grown many organisations do not realise what AI tools employees are using and the types of work being undertaken. All AI systems, underlying processes and tools need to be accountable to the individuals using them and in turn the individuals using AI for their work need to be accountable for how they are using AI. Anyone using AI in an organisation needs to be able to explain how the AI decision framework is working, how AI is interpreting data, where the data sources in the LLM are being taken from, and how AI will evolve as it becomes more sophisticated as part of their working patterns. If users cannot explain these areas, they are liable to be leaving the organisation open to trust issues in terms of how data is being used, processed and updated to make AI based decisions. IT teams need to make sure the right governance and compliance safeguards are in place for users which can be easily understood to identify and stop AI trust issues before they become problematic and difficult to manage.

Ensuring AI is reliable and performing to expectations: Time needs to be taken to understand and test the reliability of different AI tools. This should include testing the functionality and decision framework of any AI systems and processes. By doing this, outcomes which are incorrect or could cause performance issues can be identified and plans put in place to resolve. Secondly performance related to AI needs to be considered not just for the different tools being used but underlying factors such as infrastructure and security in terms of data in LLM's being used by different AI system components, algorithms need to be secured against the evolving threats of unauthorised access, corruption and attack.

Key considerations and actions to implement 'Trust by Design' into an organisations AI strategy

In terms of developing an AI strategy which is likely to touch on various stakeholder groups, Trust by Design needs to be considered at each stage and for different groups. Key questions and areas for exploration as part of Trust by Design in your organisations AI strategy should include:

Does your organisation understand why AI is different to other technology areas in terms of trust?

Most technologies are there to solve historical problems or enable organisations to solve/resolve current challenges they are facing. They do not adapt in real time to user requirements/behaviours or learn from new data inputted. AI is different. Most AI tools adapt in real time based on new data being placed into the LLM's and learns from its own use resulting in different decisions being made by the technology each day.

These changes in AI systems need to monitored, managed, validated and acted upon if they are detrimental to an organisation and can impact trust. Examples of where organisations need to review their AI tools and understand why they are different to other technology areas and the impact of Trust by Design include:

- How the historical data in LLM's is being used in different company decisions across Customer Service, HR, Sales and Marketing and Procurement before an AI tool is identified to help. By understanding the types, volumes of data and different stakeholder groups the data being processed/used and factoring in trust issues early it will resolve many of the poor user experiences associated with AI.

- Are the algorithms LLM's use for AI producing biases, reproducing historical mistakes, not aligning with the right governance processes and have the right laws and financial implications been factored into AI strategy design?

- The right AI tool may not be the right fit for the organisation based on its culture, values or services the organisation provides. This brings into play ethical and social trust issues which need to be considered.

- Who is responsible for selecting the LLM to use for AI purposes? Do they hold or have particular biases towards the data they are purchasing? Is the LLM provider trustworthy and have they undertaken trust and bias checks on the data inputted into the model?

- Is the AI development team made up of people from a range of different backgrounds/ experiences so diverse viewpoints are captured and understood before an AI tool is designed and deployed. If an AI tool is designed and deployed based on 1 or 2 people's opinions it may not be effective in it's use as it will not be considering different types of data and information based on feedback/input from a range of stakeholders not just those who are designing the tool.

- Have the right security measures around AI been considered as part of an AI tool deployment. For example, is the security risk framework in place, what will happen if LLM's are suffering from data poisoning or the LLM's are hallucinating and producing false information, can existing threat models proactively understand the real time changes AI will make when learning? These can all lead to AI performance suffering and creating trust issues around data resulting in reputational and legal issues due to trust not being embedded as part of security in AI design.

What are the trust risks when designing and building an AI strategy?

- The main risks are legal, financial and reputational. If your AI tools are not designed correctly with the right data management, security and governance in place they can be corrupted and lead to AI providing the wrong answers to different stakeholder groups. For example, customers could become disgruntled as they may have received the wrong information about a product/service which impacts their time and potentially lead to financial losses, suppliers may not be paid correctly, and employees may find their sensitive data being used in projects which they did not sanction its use, leading to legal action.

Why is building Trust by Design into an AI strategy important?

- Quite simply if humans do not trust a technology like AI to make decisions for them, they won't use it or be sceptical of using it. This could result in the technology not being fully adopted or requiring significant management where it becomes cost prohibitive for an organisation to manage.

- Many organisations need to take a step back with their AI strategy and assess their controls, policies and processes for interacting with different stakeholder groups. They need to identify where trust issues with different stakeholder groups are happening and are likely to occur in the future with AI tools deployed.

- As a starting point organisations need to analyse where the biases in data are occurring, review algorithms in terms of how machines are learning from the data it is using to make decisions and the decision making models, where the data for AI tools is being hosted (is it in a safe, sure and compliant environment), and do they have the right talent available to develop the AI tools in the organisation to make sure trust is embedded in each AI tools being used. After these areas have been considered an evaluation needs to take place against governance, risk and compliance frameworks, before a roadmap is put in place to rectify the potential trust issues.

How can the right roadmap be put in place to mitigate trust risk in the AI strategy?

Identifying and understanding where trust issues may occur as part of AI is only the first part of solving the potential challenges. To overcome the challenges the right trust roadmap needs to be put in place. This should include:

- Understanding the different types of AI tools available for your organisations requirements.
- Assessing different AI Tools vs organisational priorities and the potential impact on trust.
- Creating clear objectives with indicators where Trust will play an important part in any AI strategy deployed and assessing the underlying processes, tools and LLM's.
- Identifying the technology infrastructure you will need to make your AI strategy a reality and what improvements need to be made to improve trustworthiness.
- Building an AI roadmap which has Trust by Design factored into it with established Ethical and Governance guidelines.
- Hosting regular review and improvement Trust by Design workshops so as AI evolves in the organisation changes in user, customers and other stakeholder groups sentiment can be captured and acted upon.

How can your organisation maintain the right trust standards when implementing AI?

Trust needs to embedded in in an AI strategy from the beginning. Organisations need to understand how different AI decisions are being made and used. Then review each decision being made against a use case. To do this effectively it is recommended:

- The organisation undertakes an inventory of the different AI tools being used across the organisation and the LLM's underpinning each AI system.

- Set up an AI ethics and trust board to review the different AI systems being used and use the board members with their experience of ethics, law, data privacy, compliance regulation to assess where AI tools may cause trust issues.

- Put in place a set of AI design standards for developing AI tools in your organisation. They should not only look at the technology aspects of AI but should include governance and accountability standards and draw on people's experiences of where AI could cause harm to the organisation.

- Create a regular AI inventory and impact trust assessment. This will enable a regular review of AI tools being used and highlight potential trust issues which need to be rectified against an AI trust framework.

- Use validation tools to assess in real time if the algorithms being used are performing to expectations and producing accurate, fair and unbiased outcomes. The validation tools used should also be mature enough to track any changes AI is making and alert relevant managers to act if the decision being made by an AI tool could impact the company reputationally, legally or financially.

- Use an NCSC assured consultancy partner to undertake an independent security audit of AI tools. They should be able to assess the resilience of infrastructure to cope with AI, identify where data breaches are occurring and could impact trust both today and in the future.

- Finally and probably most important is providing users of AI tools with the right training on trust, ethical and social guidelines. AI tools are no longer just used by the IT department. They are being used by many different user groups across the organisation. Many of the users of AI tools will only be aware of how AI tools are helping them be more productive in their work and not the key trust issues they may be impacting by using a particular AI tool.

Final thoughts

In this article we have discussed why Trust by Design is a crucial consideration when designing an organisations AI strategy. AI strategy development is a collective responsibility where trust needs to be embedded at every stage from initial assessment of where AI can help an organisation to resolve a problem to making sure data sourced is secure and trustworthy, algorithms are designed correctly, and infrastructure can cope with changing AI requirements in real time.

Achieving a trusted status with AI is not an easy task. But those organisations who start to embed Trust by Design into their AI strategies will be better placed to improve their customer, supplier and stakeholder relationships both now and in the future. They will be able based on different stakeholder groups having trust in their AI strategy, be able to take advantage of this technology and unlock the true potential of AI for their organisation.

[1] https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/articles/understandingaiuptakeandsentimentamongpeopleandbusinessesintheuk/june2023

[2] https://www.gartner.com/en/newsroom/press-releases/2023-07-27-gartner-survey-finds-55-of-organizations-that-have-deployed-ai-take-an-ai-first-strategy-with-new-use-cases

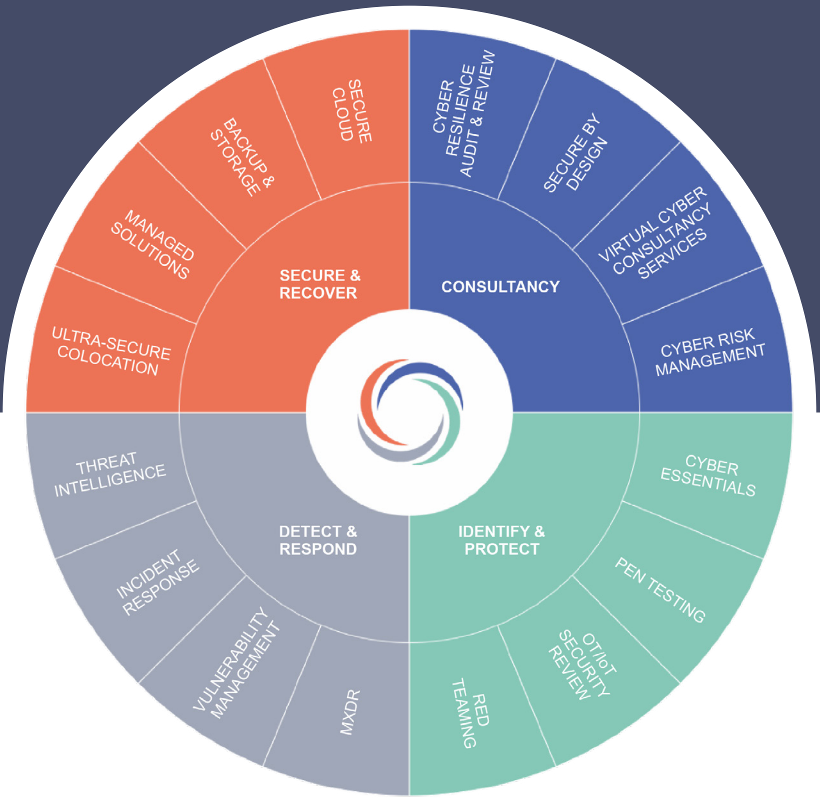
[3] https://www.trade.gov/market-intelligence/united-kingdom-artificial-intelligence-market-2023

Discover more about Cyberfort's all-encompassing Cyber Security Services

At Cyberfort we provide a range of customers with all-encompassing Cyber Security Services. We are passionate about the cyber security services we deliver for our customers which keeps their people, data, systems and technology infrastructure secure, resilient and compliant.

Our business offers National Cyber Security Centre assured Consultancy services, Identification and Protection against cyber-attacks, proactive Detection and Response to security incidents through our security operations centre and a Secure and Recover set of Cloud solutions which keeps data safely stored, managed and available 24/7/365.

Over the past 20 years we have combined our market leading accreditations, peerless cyber security expertise, strong technology partnerships, investment in our future cyber professionals and secure locations to deliver a cyber security experience for customers which enables them to achieve their business and technology goals in an ever-changing digital world.



For more information about our Cyber Security services please contact us at the details below:

+44 (0)1304 814800 | info@cyberfortgroup.com | https://cyberfortgroup.com

We look forward to working with you