

Analysis on Security Aspects of the Framework for Generative AI in Dŵr Cymru Welsh Water

BSc (Hons) Computing (Computer Networks and Cybersecurity) Kes Sumner

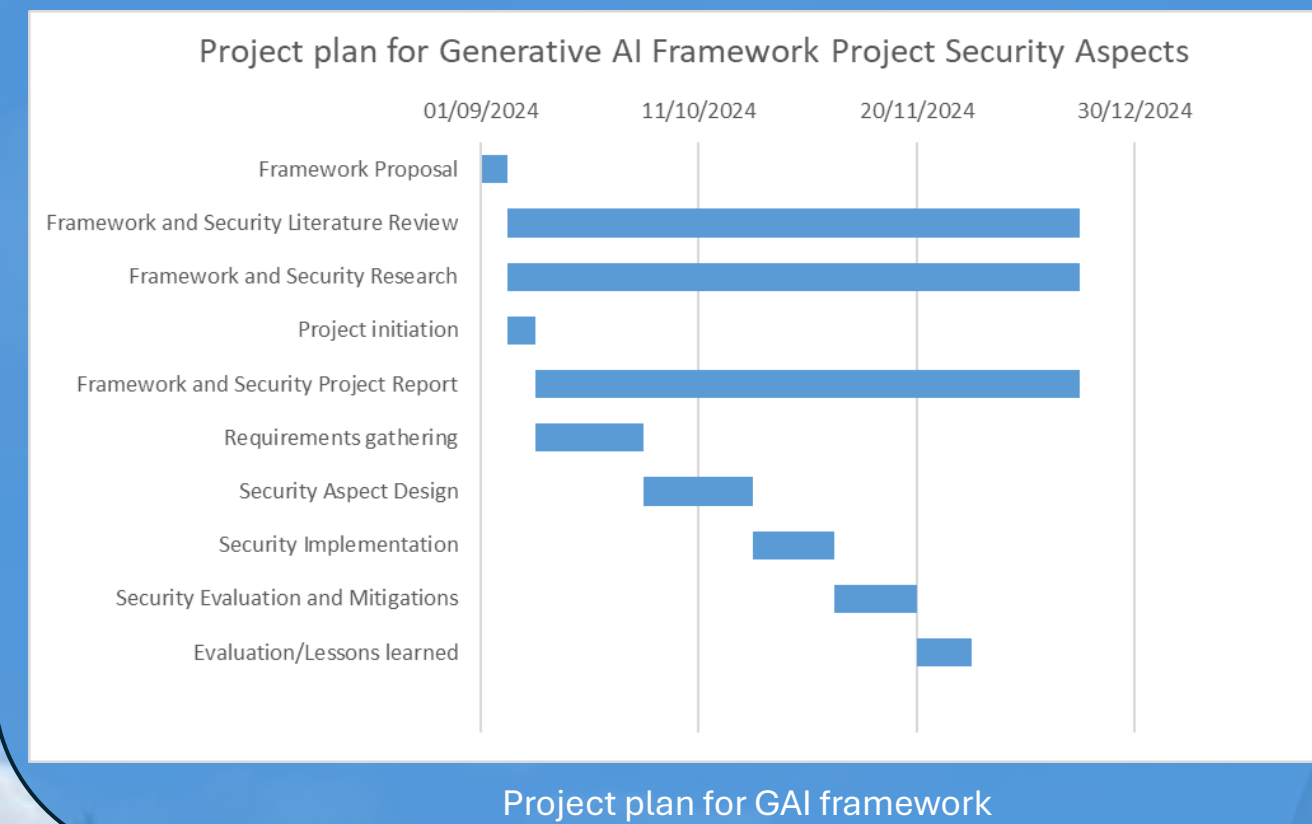
1. Conclusion

The project stages encompassed the research for the creation of a new framework for the secure and safe use of generative artificial intelligence (GAI) Dŵr Cymru Welsh Water (DCWW).

The literary review and research identified several key aspects of risk, concern and consequence to be included in the new framework. This framework will enable secure, safe and competent usage of GAI within the organisation.

Security is a critical aspect of the framework. A selection of cyber security risks were identified to be included in a few designed tests. Demonstrations of successful mitigations of the attacks were carried out, with a focus on the relevance of the security risks to a GAI system.

4. Planning



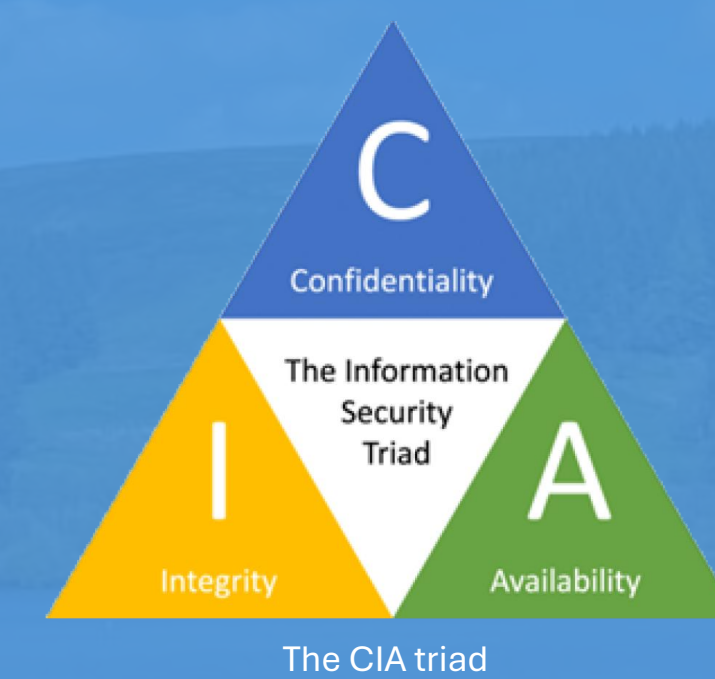
The project contingency plan for designing, implementing and evaluating the security aspects of the framework follows a waterfall model rather than an agile model [1].

6. Security Risks

Several of the cyber security risks included within the framework were demonstrated in a virtual environment and successfully mitigated as part of the evaluation phase of the project. The select risks were:

- Access Control
- Phishing
- Session Hijacking

Generative AI systems can pose serious security threats, such as the above, that undermine all three points of the Confidentiality, Integrity, Availability (CIA) triad. The CIA triad is a globally recognised cornerstone information security model [2], consisting of three fundamental principles.



- Confidentiality – “the fact of private information being kept secret” [3]
- Integrity – “Guarding against improper information modification” [4]
- Availability – “Ensuring timely and reliable access to and use of information” [5]

2. Project Aim and Objectives

Aim:

Create and evaluate a draft framework for GAI usage within DCWW, demonstrate security aspects and report on the outcome.

Objectives:

Carry out a critical literature review, focusing on security and risk around GAI and the mitigations required for a generative AI framework.

Perform appropriate research, consider and adapt to security and ethical considerations and any relevant regulatory requirements.

Design and build a draft framework for the use of GAI in DCWW.

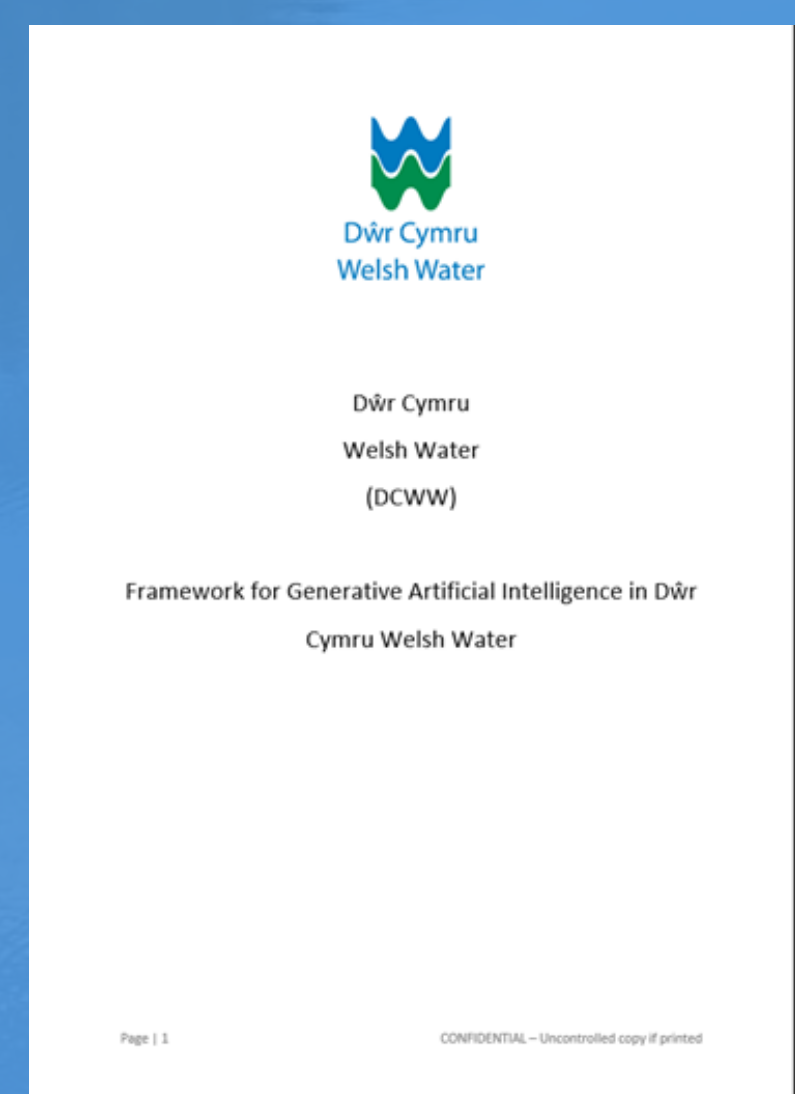
Identify, evaluate and mitigate various key related cyber security risks.

5. Framework Structure

Six key aspects of GAI were identified as areas of risk, concern or consequences:

- Security – prompt injection, access control, phishing for example.
- Ethics – bias and discrimination and unintended consequences.
- Accuracy and reliability – hallucinations, misinformation.
- Legal – intellectual property and ownership of GAI generated content.
- Regulatory – data privacy, transparency and explainability.
- Reputation – damage to reputation from consequences of the above risks.

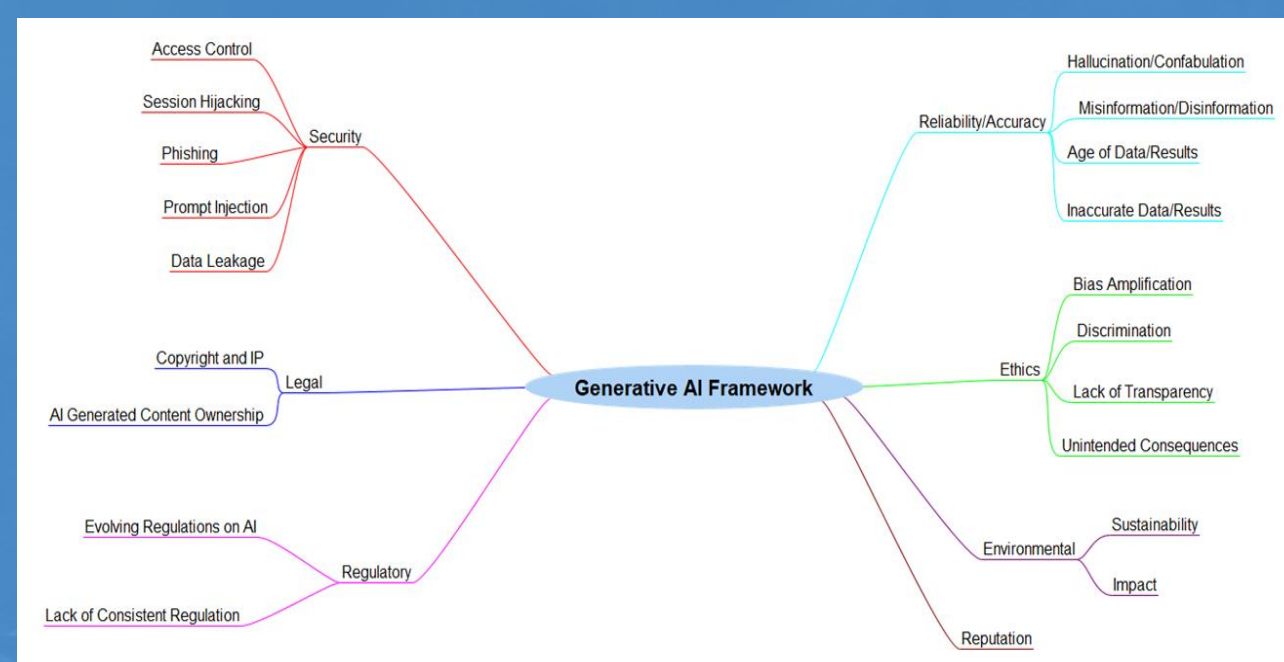
These aspect formed the core structure of the framework document and contributed to the developed the guideline statements within the framework:



1	Introduction	5
1.1	Background	5
2	Purpose	6
3	Scope	8
4	Framework	9
4.1	Cybersecurity	10
4.1.1	Cybersecurity risks/threats	10
4.1.2	Guideline statements	11
4.2	Ethics	11
4.2.1	Ethical risks and concerns	12
4.2.2	Guideline statements	13
4.3	Accuracy and reliability	13
4.3.1	Accuracy and reliability risks	13
4.3.2	Guideline statements	13
4.4	Legal Implications	13
4.4.1	Intellectual Property and Copyright Issues	14
4.4.2	Guideline statements	14
4.5	Regulatory and Compliance	14
4.5.1	Compliance and Regulatory Issues	14
4.5.2	Guideline statements	15
4.6	Reputational Impact	15
4.6.1	Reputational Damage	15
4.6.2	Guideline statements	15
5	Framework Guidelines	16
6	DCWW Compliance	17
7	Related Documentation	17
8	Responsibilities	17
9	Definitions	18
10	References	19
11	Appendices	20
11.1	How GAI works	20
11.2	AI Steering Group	20

Framework document structure

3. Literature Review



Mind map of GAI risks

Initial high-level research and use of some generative AI platforms helped develop a mind map of the key areas of security risk and other concerns, such as ethical and legal issues.

7. References

- [1] T. Thesing, C. Feldmann and M. Burchardt, "Agile versus Waterfall Project Management: Decision Model for Selecting the Appropriate Approach to a Project," *Procedia computer science*, vol. 181, pp.746-756, 2021. doi: 10.1016/j.procs.2021.01.227
- [2] J. Nikander, O. Manninen and M. Laajalahti, "Requirements for cybersecurity in agricultural communication networks," *Comput.Electron.Agric.*, vol. 179, pp.105776, 2020. doi: 10.1016/j.compag.2020.105776.
- [3] J. Rehberger, "Cambridge Online Dictionary," [Online]. Available: <https://dictionary.cambridge.org/dictionary/english/confidentiality>. [Accessed Jan. 04, 2025].
- [4] J.Cawthra et al, "Data Integrity: Identifying and Protecting Assets Against Ransomware and Other Destructive Events," NIST, Dec. 2020. [Online]. Available: <https://csrc.nist.gov/pubs/sp/1800/25/final>. [Accessed Jan. 04, 2025].
- [5] NIST, "Computer Security Resource Center," [Online]. Available: <https://csrc.nist.gov/glossary/term/availability>. [Accessed Jan. 04, 2025].