

Practical guide to Artificial Intelligence risks and mitigations for Trusted Research Environments and the RELEASE-AI framework



University of Dundee

and the RELEASE-AI framework

Alba Crespi-Boixader^{1,2,*}, Simon Li^{1,2}, James Liley³, Laura Ward^{1,2}, Christian Cole^{1,2}, Jim Smith⁴



Durham University



¹Division of Population Health and Genomics, ²Health Informatics Centre – University of Dundee, ³University of Durham, ⁴University West of England

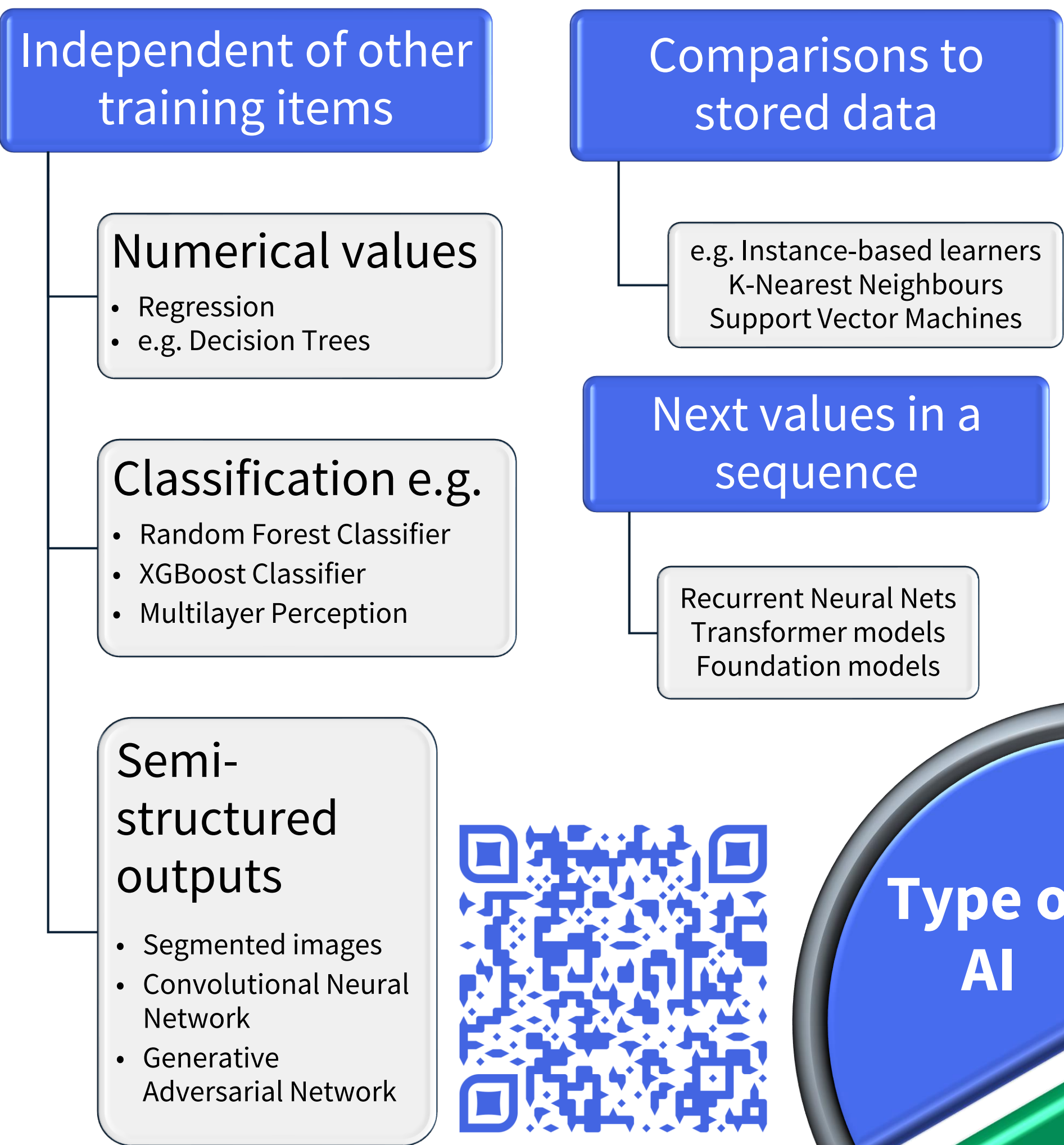
*acrespi001@dundee.ac.uk

Motivation

- As the use of Artificial Intelligence (AI) and Machine Learning (ML) becomes increasingly common in research, ensuring responsible and secure handling of sensitive data is essential—particularly within Trusted Research Environments (TREs), as stated in GRAIMATTER[1].
- Need to clearly identify which data privacy risks TREs would face, and these will depend on the type the user or researcher employs.
- AI/ML projects present unique challenges to disclosure control, and they require procedures to be adapted and expanded to address the specific risks associated with model development and deployment.
- Risks can be minimised and avoided in most cases but cannot always be completely eliminated.
- Specialised tools, like SACRO-ML[2], should be used prior to model release.

Type of AI/ML

Taxonomy based on the outputs produced.



Disclosure risks

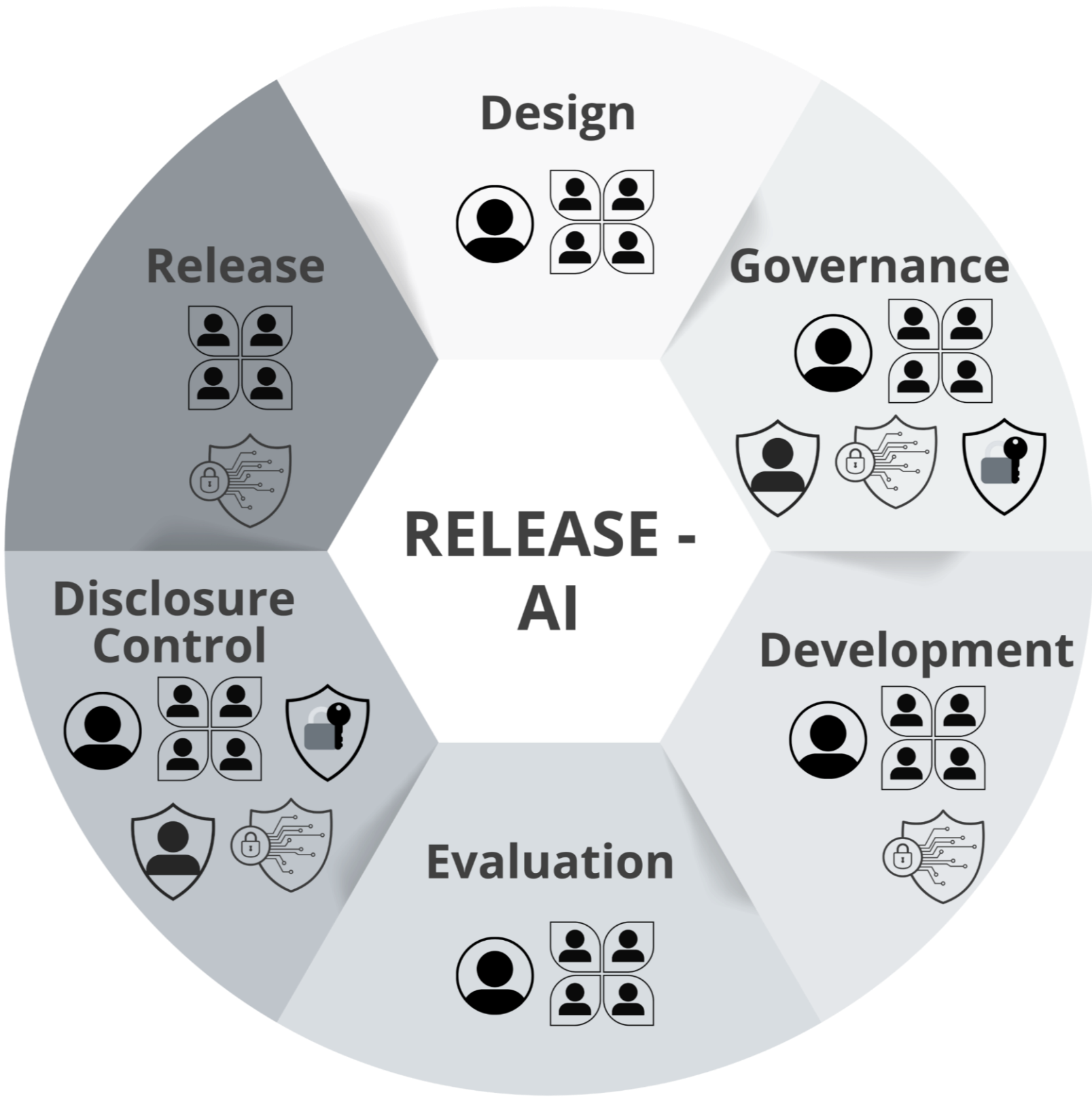
Privacy risks associated with the types of AI/ML.



RELEASE-AI framework

Risk Evaluation and Lifecycle Accountability for Secure Egress of Artificial Intelligence

- RELEASE-AI is a TRE framework for AI/ML projects across six phases of the lifecycle with sensitive data.
- RELEASE-AI provides stakeholders with clear recommendations on measures to ensure compliant and successful AI/ML projects in TREs.
- RELEASE-AI promotes early identification of potential risks with corresponding mitigation.
- RELEASE-AI assigns roles and responsibilities throughout the project.



- Researcher/ TRE User
- Project Team
- Output Checkers
- TRE Operator
- Data Controller



Acknowledgments

This work was funded by UK Research & Innovation [Grant Number MC_PC_24038] as part of Phase 2 of the DARE UK (Data and Analytics Research Environments UK) programme. DARE UK is funded by UK Research and Innovation (UKRI) and led by Health Data Research UK (HDR UK) and ADR UK (Administrative Data Research UK).



References

- [1] E. Jefferson et al., doi:10.5281/zenodo.7089491
- [2] J. Smith, et al., arXiv:2212.01233