APPENDIX 8: Decision tree for complex synthetic data scenarios

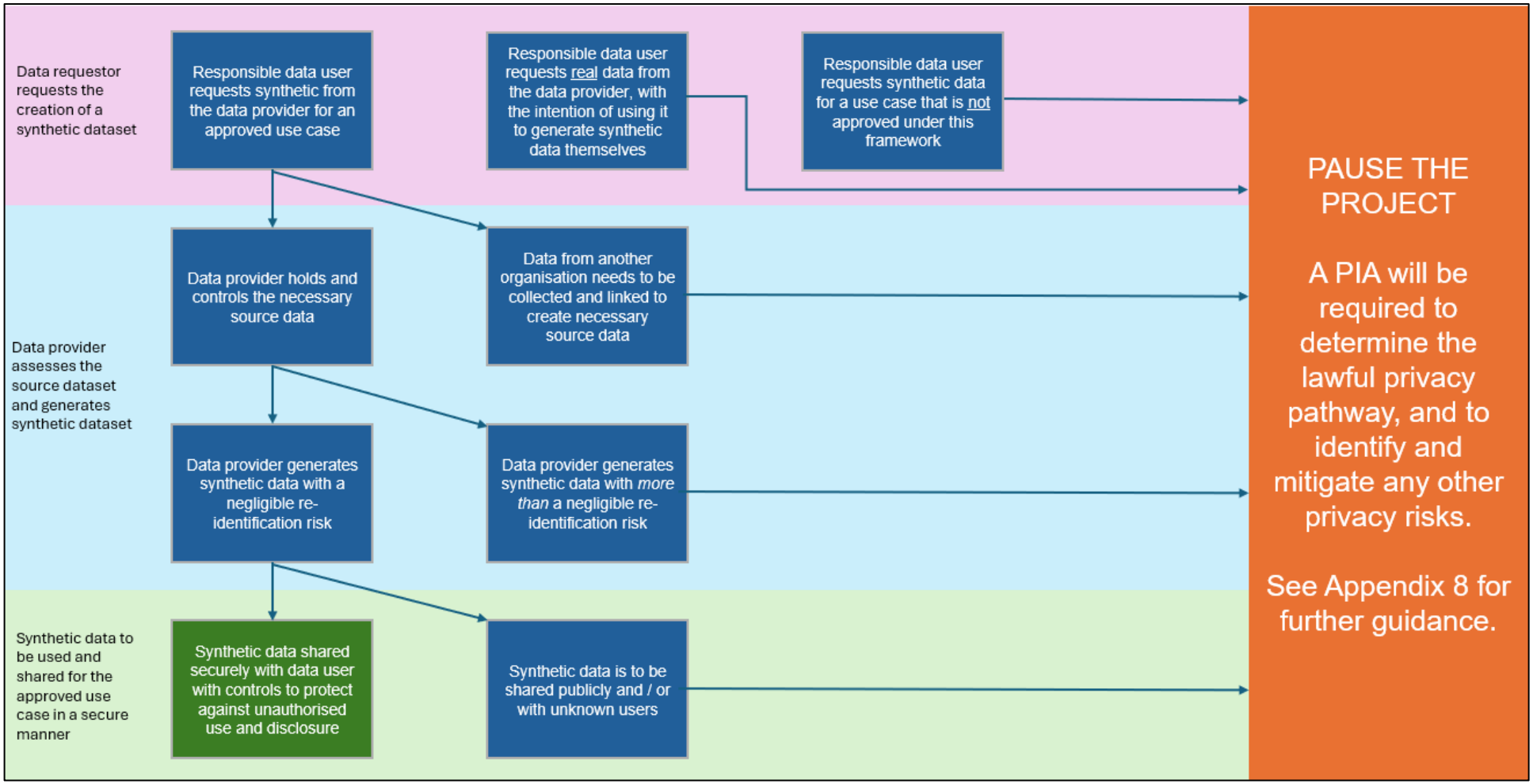
The five-step approach for generating and using synthetic data outlined in this framework should be sufficient for managing privacy risks where a synthetic data request involves a fairly straightforward workflow.

A sample scenario of a straightforward workflow would be as follows:

1. A SynD organisation (e.g. the University of Sydney as the Data Requestor) requests another SynD organisation (e.g. NT Health as the Data Provider) to generate a synthetic health dataset using source data already held by that Data Provider, for an acceptable use case.
2. The source data is fit for purpose, and the resulting synthetic dataset is considered robustly and effectively de-identified with only a very low risk of re-identification.
3. The Data Provider provides the Data Requestor with access to the synthetic dataset in a secure manner and continues to have an appropriate level of oversight of the handling of the dataset.

However, for scenarios that involve more complex workflows, *or* where any of the steps in this framework cannot be successfully completed, material privacy and legal compliance risks may emerge. In these cases, additional steps are needed to manage these risks.

The decision tree below anticipates some of these more complex scenarios where additional steps and assessments are required, to ensure the synthetic data request can proceed lawfully.


Data requestor requests the creation of a synthetic dataset
Responsible data user requests synthetic from the data provider for an approved use case
Responsible data user requests real data from the data provider, with the intention of using it to generate synthetic data themselves
Responsible data user requests synthetic data for a use case that is not approved under this framework
Data provider assesses the source dataset and generates synthetic dataset
Data provider holds and controls the necessary source data
Data from another organisation needs to be collected and linked to create necessary source data
Data provider generates synthetic data with a negligible re-identification risk
Data provider generates synthetic data with more than a negligible re-identification risk
Synthetic data to be used and shared for the approved use case in a secure manner
Synthetic data shared securely with data user with controls to protect against unauthorised use and disclosure
Synthetic data is to be shared publicly and / or with unknown users
PAUSE THE PROJECT
A PIA will be required to determine the lawful privacy pathway, and to identify and mitigate any other privacy risks.
See Appendix 8 for further guidance.