Halderstone

Training module

AI Limitations & Failure Modes

AI uncertainty, limitations and common failure modes across predictive and generative AI systems

Do you understand how AI systems can fail?

Overview

AI systems operate under uncertainty. Their behaviour shifts with data, context, integration choices, and human interaction. When these limits are not understood, organisations either over-trust or over-block AI, creating governance and operational risk.

This module develops a clear, structured view of AI limitations and failure modes across predictive and generative systems. Participants learn where uncertainty originates, how failures emerge in real environments, and how to interpret AI outputs and technical evidence with appropriate judgment.

Target audience

- AI management system managers and implementers working with technical teams
- Governance, risk, and compliance professionals who need AI domain fluency
- Product owners and process owners responsible for AI-enabled services
- Auditors who need a shared baseline understanding of AI systems (not audit craft)
- Anyone who wants to get a basic understanding of AI fundamentals

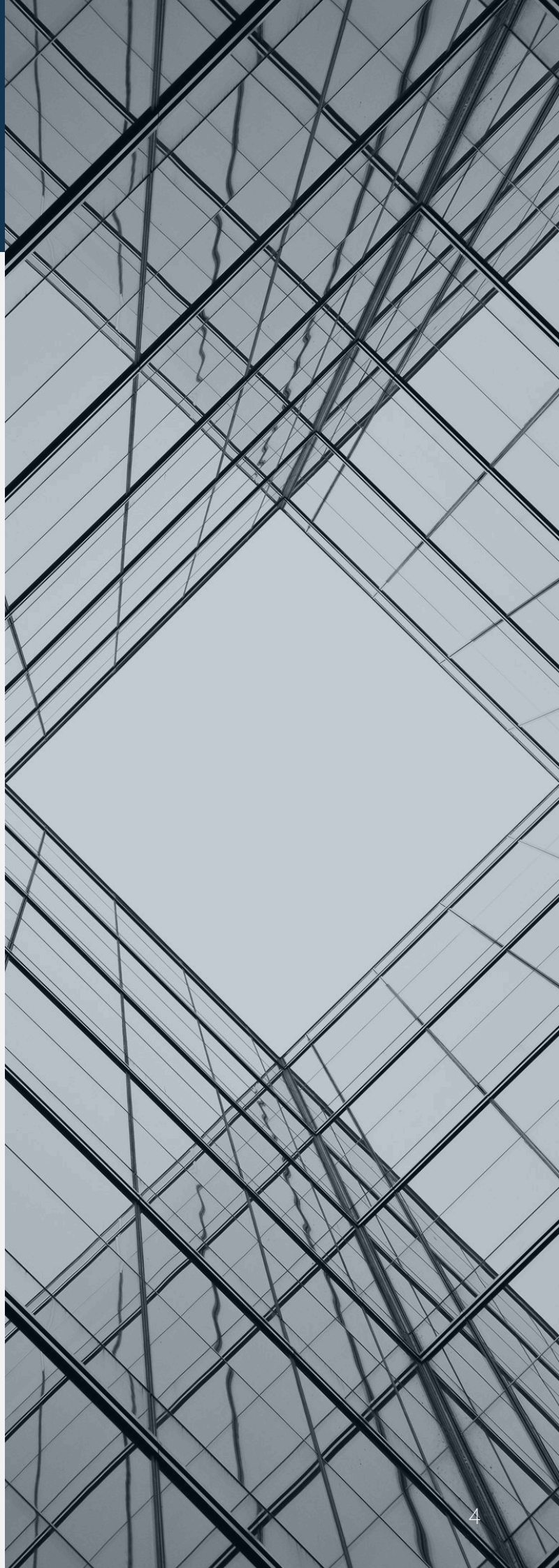
Is this module for you?

It is a good fit for you if you...

- want to understand why AI outputs should be treated as fallible signals rather than as verified facts.
- need a realistic mental model of uncertainty in predictive and generative AI.
- want to recognise common AI failure modes in real operational contexts.
- need to reason about reliability, limits, and confidence without false certainty.
- want to interpret AI behaviour without relying on vendor claims or tooling.

It may be less suitable for you if you...

- are looking for AI risk assessment methods or control design.
- expect statistical deep dives or model-level optimisation techniques.
- want implementation playbooks, monitoring setups, or lifecycle processes.
- already have advanced AI research or data science expertise.



Learning outcomes

Key outcomes

- Explain the main sources of uncertainty in AI systems and how they affect outcomes
- Recognise common failure modes in predictive machine-learning models and generative AI systems
- Describe socio-technical failure modes where human and organisational factors interact with AI

Additional capabilities

- Identify data pipeline issues that contribute to model drift and bias
- Conduct structured failure-mode walkthroughs to anticipate how AI performance may degrade
- Communicate limitations and uncertainty to stakeholders to support responsible AI use

Agenda

Why limitations and uncertainty are central to AI governance

How AI outputs should be treated as fallible probabilistic signals whose content is not automatically reliable or factually correct, and how organisations typically fail by either over-trusting or over-blocking AI systems

Where uncertainty comes from in AI systems

How uncertainty arises from data, labels, and context shifts, and how model uncertainty differs from system-level uncertainty caused by integration, workflows, and human factors

Model behaviour limits in predictive ML

How predictive models are constrained by generalisation limits, spurious correlations, distribution shift, and performance trade-offs that depend on thresholds and use context

Model behaviour limits in generative AI

How generative models exhibit hallucinations, instruction-following limits, prompt sensitivity, and constraints related to context windows and knowledge updates

Data-related failure modes

How failures emerge from training–serving skew, data leakage, silent pipeline changes, and data quality issues such as missingness, bias, and weak provenance

System and socio-technical failure modes

How AI systems fail through automation bias, misuse, feedback loops, gaming, and brittle interfaces affected by latency, integration constraints, and fallback behaviour

Case-based workshop

Applying the learned concepts, methods, and approaches in a realistic case setting

Included materials

Learning materials

- Slide deck
- Participant workbook

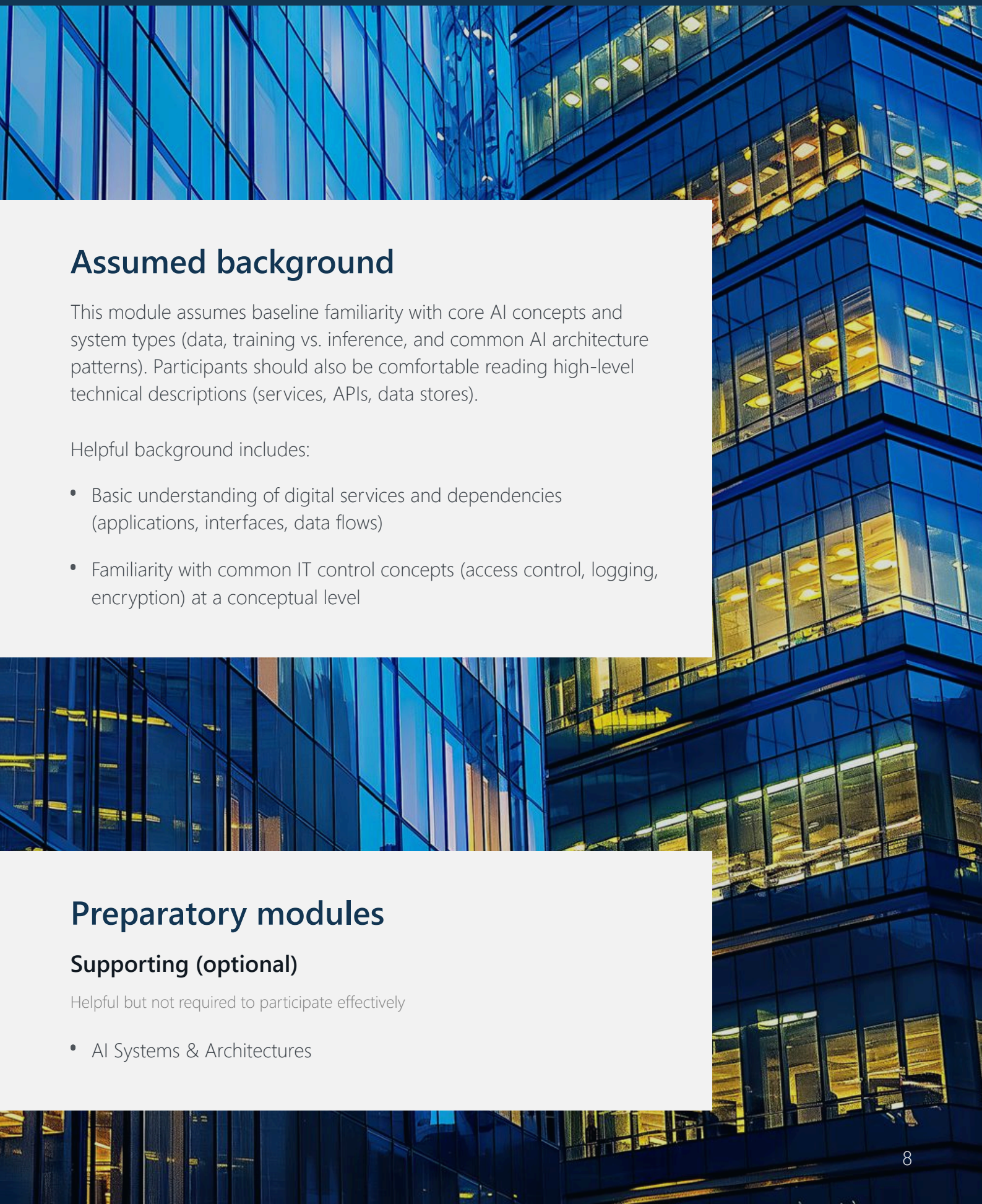
Templates & tools

- AI uncertainty map (sources and observable signals)
- Failure mode catalogue (predictive, generative, and system-level)
- Case walkthrough canvas (data → model → integration → use)
- Evidence question set for technical walkthroughs

Confirmation

- Confirmation of participation

Preparation guidance



Assumed background

This module assumes baseline familiarity with core AI concepts and system types (data, training vs. inference, and common AI architecture patterns). Participants should also be comfortable reading high-level technical descriptions (services, APIs, data stores).

Helpful background includes:

- Basic understanding of digital services and dependencies (applications, interfaces, data flows)
- Familiarity with common IT control concepts (access control, logging, encryption) at a conceptual level

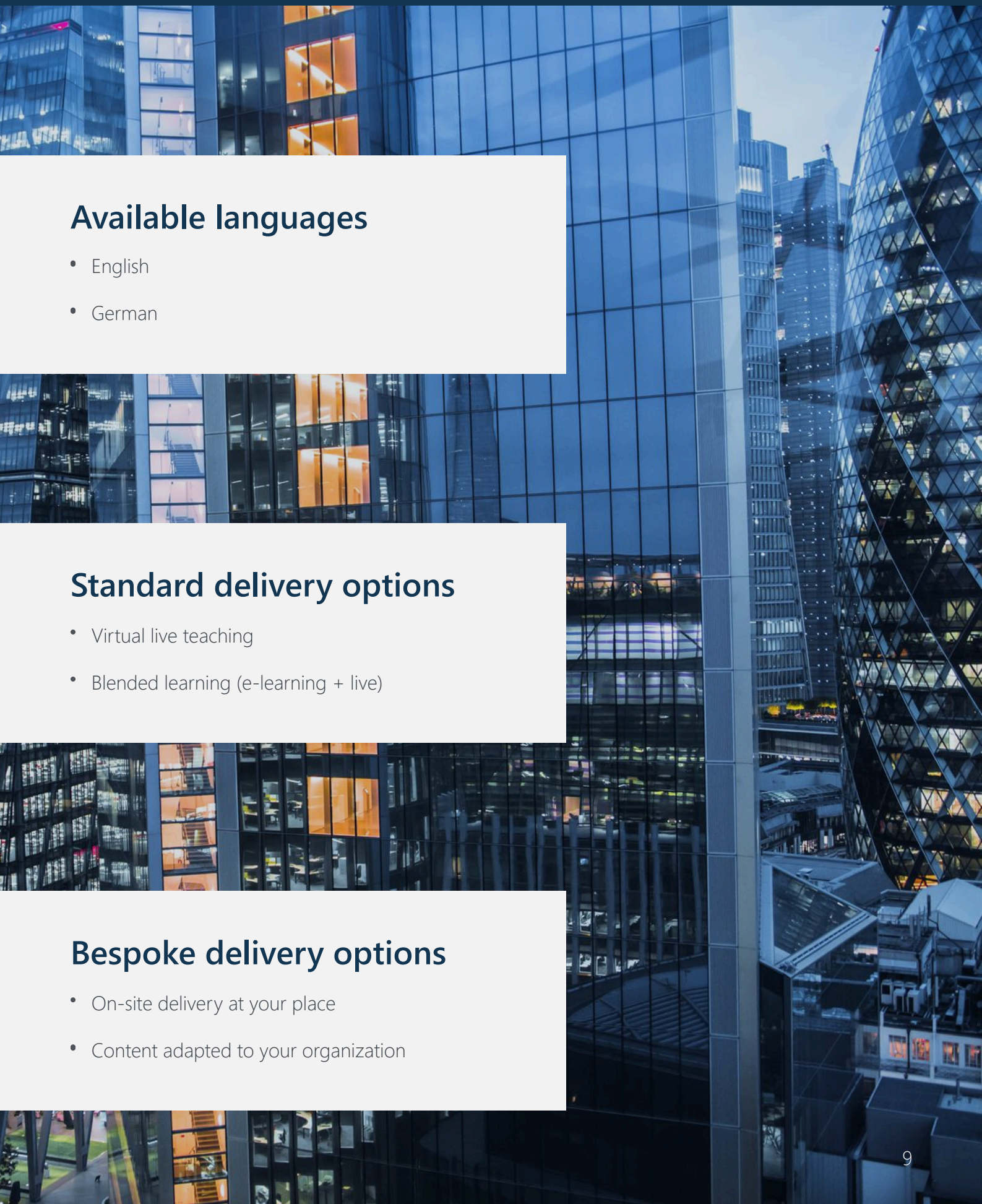
Preparatory modules

Supporting (optional)

Helpful but not required to participate effectively

- AI Systems & Architectures

Logistics



Available languages

- English
- German

Standard delivery options

- Virtual live teaching
- Blended learning (e-learning + live)

Bespoke delivery options

- On-site delivery at your place
- Content adapted to your organization

Halderstone

Halderstone by Langer & Co
Zürcherstrasse 2
CH-8852 Altendorf
Switzerland
info@halderstone.com
www.halderstone.com