AI and the Future of Threat-Led Penetration Testing

Capability, control and credible assurance

November 2026

CREST

Contents

1. Introduction and research methodology	3
2. Executive Summary	4
3. AI is increasingly used in TLPT, but cautiously	5
4. AI can shape the test before execution begins	8
5. More analytical capacity raises the bar for threat fidelity	10
6. How AI authority changes the control model	12
7. Assuring more than just the model	16
8. As capability becomes easier to access, judgement becomes more valuable	20
9. Controls can make a test safer while changing what it proves	22
10. Acceptance and scrutiny are increasing together	24
11. Conclusion	27

1. Introduction

Threat-led penetration testing (TLPT) connects intelligence about relevant adversaries with a controlled examination of an organisation's security and resilience. AI is changing the speed, scale and shape of that work. It can influence which intelligence is prioritised, which scenario is selected, how tooling is created, the pace of technical activity and how findings are interpreted. As adversaries make greater use of AI, TLPT teams need to consider whether an engagement can credibly represent those capabilities and what level of AI-enabled capability is required within the test itself.

CREST has brought together a provider survey and two industry roundtables to examine those changes. The survey provides the primary evidence on reported adoption, use, skills and expectations across commercial (i.e. engagements that are not part of a regulatory requirement or framework) and regulated or scheme-based delivery. The roundtables add practitioner context, particularly from financial services, and help test what the survey findings could mean for engagement design, professional judgement and assurance in practice.

The research surfaces several tensions that matter for the future of TLPT. Greater AI capability can widen the range of techniques available while raising the bar for assurance. More realistic techniques can introduce additional risk. Human control improves safety and can also alter the conditions the test is intended to reproduce. Easier access to sophisticated tooling may shift the value of external providers towards judgement, scenario design and independent assurance. This report examines those tensions and what they mean for providers, practitioners, buyers and assurance stakeholders.

Research Methodology

A company-level survey gathered responses from 26 TLPT providers: 14 in the commercial group and 12 in the regulated group. They include specialist providers, global consultancies and large technology organisations, headquartered across approximately 10 countries and operating internationally.

CREST also convened 2 industry roundtables involving highly experienced providers and institutional stakeholders, with a particular emphasis on financial services. The discussions explored emulation, risk, capability development and community support, drawing out practical concerns, areas of alignment and questions requiring further work. These insights have not been directly attributed and inform the analysis throughout this report.

Throughout the report, 'respondents' refers to survey answers and 'participants' to roundtable contributions. Percentages relate only to the survey; the roundtable discussions provide qualitative insight, not additional survey responses or independent validation of the results.

2. Executive Summary

AI use is becoming routine within threat-led penetration testing (TLPT). Most respondents already use it beyond research-only activity, generally at defined points within practitioner-led services. The emerging model is selective integration, with practitioners retaining accountability for the engagement.

The survey shows that 21 of 26 respondents use AI beyond experimentation, while 2 report extensive integration across the lifecycle. This pattern makes the way AI is introduced into the service more important than adoption alone. Providers need to understand where AI supports the work, what information it can access, how its outputs are reviewed and which activities remain subject to practitioner approval.

AI is also expanding the number of options available to a testing team. Among respondents using AI for threat actor emulation, 86% report improved accuracy and realism. Greater analytical capacity raises the bar for judgement because teams still need to decide which generated insights, attack paths and techniques are credible for the relevant actor and useful for the assurance objective. The quality of the test depends on choosing the right threat and maintaining the connection between intelligence, scenario and activity.

Authority is becoming a central assurance question. Human decision points remain prominent, especially around execution, and no respondent carrying out regulated engagements reports autonomous execution. The key boundary is what AI can access, influence or do before a practitioner intervenes, supported by monitoring, stop conditions and evidence of material decisions.

The same logic applies to the technology itself. Two providers can use the same underlying model and present very different assurance profiles because much of the capability may sit in integrations, permissions, data flows, orchestration and human control. Model identity on its own provides limited insight into the assurance of an AI-enabled TLPT service.

Professional judgement becomes more valuable as AI makes more analysis and technical capability available. Validation of AI-generated technical output is rated essential by 92% of respondents, while manual exploitation, secure data handling and understanding AI limitations also rank highly. Effective use depends on the competence to assess, challenge and stand behind generated outputs.

The market may also change on both sides of the engagement. Providers expect greater use of AI-generated offensive tooling, stronger demand for AI-specific skills and more procurement, disclosure and audit requirements. Some financial institutions are also building more capability themselves. As access to technology becomes less distinctive, independent challenge, threat relevance, complex scenario design and credible assurance may become more important sources of provider value.

Taken together, the findings show AI changing where decisions are made, what needs to be evidenced, how safety and realism are balanced and what buyers should expect from a credible TLPT service. Assurance increasingly needs to focus on what AI changes about the engagement and the confidence that can be placed in the result.

3. AI is increasingly used in TLPT, but cautiously

AI is supporting a growing share of TLPT work, but the prevailing model remains selective integration within practitioner-led services. Providers are introducing AI into particular activities and defined parts of delivery rather than replacing the methodology as a whole.

Operational adoption is becoming normal, not uniform
21 of the 26 respondents report AI use beyond experimentation. 19 (73%) describe selective live use or integration within defined parts of their methodology, while 2 report extensive integration across the lifecycle. The result shows AI becoming operational across the cohort without pointing to a single model of adoption.

Table 1. Current AI use by primary TLPT delivery context

Reported use	All responses	Commercial	Regulated
Experimental or research use only	5 (19%)	4 (29%)	1 (8%)
Selective live use	10 (38%)	5 (36%)	5 (42%)
Embedded within defined parts of the methodology	9 (35%)	3 (21%)	6 (50%)
Extensively embedded across the TLPT lifecycle	2 (8%)	2 (14%)	0 (0%)

That changes the baseline for service design. AI can no longer be treated only as an experimental exception for a small number of providers: organisations adopting it need to know where it enters delivery, what information it can use, what it is allowed to influence and how practitioners validate or authorise its contribution. Roundtable examples spanning research, threat-intelligence analysis, reconnaissance, tooling and reporting show how varied those entry points can be.

Different contexts are adopting AI differently

Half of the regulated group reports AI embedded within defined methodology, compared with 3 of 14 commercial responses. Commercial respondents are more dispersed across experimentation, selective use and extensive integration. These small-group differences show different delivery choices; they do not establish that one context is inherently more mature than another.

Roundtable participants also described AI-assisted and agentic testing in some financial institutions outside the regulated TLPT setting under discussion. These examples reinforce a broader point: an adoption claim is only meaningful when it explains the activity involved, the environment in which it occurs and the authority given to the system.

AI is supporting more of the work

22 respondents selected a higher current AI-assistance band than the band they recalled for 12 months earlier, while 4 selected the same band. 13 recalled no AI-assisted TLPT activity a year earlier; only 1 of those respondents continued to report no AI use at the time of the survey.

Table 2. Share of TLPT tasks supported by AI

Share of tasks	12 months earlier (recalled)	At the survey (current)	12 months later (expected)
0%	13 (50%)	1 (4%)	0 (0%)
1-20%	8 (31%)	8 (31%)	3 (12%)
20-40%	4 (15%)	8 (31%)	6 (23%)
40-60%	1 (4%)	7 (27%)	7 (27%)
60-80%	0 (0%)	2 (8%)	6 (23%)
80-100%	0 (0%)	0 (0%)	4 (15%)

For the following year, 17 respondents (65%) selected the 40-60% assistance band or a higher band, including 10 expecting AI to support 60% or more of TLPT tasks. These figures point towards more AI-supported work, but they do not show a corresponding transfer of decision-making or execution authority.

Assistance and autonomy are developing separately

The research separates three questions that are often grouped together: how much work AI supports, where it sits in the methodology and how much authority it receives. A provider can increase AI assistance while preserving established practitioner decision points. Assurance therefore needs to consider the scale of AI use, where it is introduced and what authority it receives.

The next section follows that influence across the lifecycle, from planning and threat intelligence through technical activity to the final assurance conclusion.

4. AI can shape the test before execution begins

AI use is activity-specific. Live use is reported most often in reporting and assessment, followed by reconnaissance and threat-informed preparation, with different patterns across planning, execution and post-execution analysis. The important point is that AI can materially influence an engagement before any AI-enabled action reaches a target system.

This early-stage use matters because planning and scoping establish what the engagement will focus on. AI is already supporting this activity for 43% of commercial respondents and 25% of regulated respondents. Where AI helps prioritise intelligence, identify areas of interest or shape an initial scenario, its influence can carry through the rest of the engagement. The data and patterns informing those outputs therefore matter: model or training-data bias could steer attention towards familiar scenarios or behaviours at the expense of organisation-specific context.

Table 3. Reported live AI use across delivery activities

Survey activity	All responses	Commercial	Regulated
Planning and scoping	9/26 (35%)	6/14 (43%)	3/12 (25%)
Reconnaissance and threat-informed preparation	13/26 (50%)	7/14 (50%)	6/12 (50%)
Execution	8/25 (32%)	4/14 (29%)	4/11 (36%)
Post-execution analysis	7/25 (28%)	5/14 (36%)	2/11 (18%)
Reporting and assessment	14/25 (56%)	6/14 (43%)	8/11 (73%)

Note: Planning and reconnaissance use a base of n=26. Execution, post-execution analysis and reporting use n=25, with one regulated response omitted at those stages.

Reporting and assessment shows a marked difference between delivery contexts: 8 of 11 regulated responses report live AI use, compared with 6 of 14 commercial responses. Commercial respondents report greater live use in planning and post-execution analysis. The pattern suggests that different contexts are choosing different points at which AI can influence the service.

That matters because influence can accumulate across the lifecycle. AI-supported planning or intelligence can shape the scenario selected for testing; that scenario can shape technical activity; and AI-supported analysis can shape the conclusions a client receives. Assurance therefore needs to consider AI influence from the point at which the engagement is scoped and the threat scenario begins to take shape.

Influence starts before execution

Initiation establishes the intended outcome and boundaries. Threat intelligence identifies relevant adversaries and behaviours. Penetration testing examines those scenarios through authorised activity, and closure brings together validated findings and lessons. AI can contribute at each point, creating a chain of influence long before or after technical execution itself.

The survey category 'reconnaissance and threat-informed preparation' captures only part of that chain. Wider threat-intelligence work includes collection, processing, contextual interpretation and scenario development. An AI-supported output at one of those stages can therefore affect later technical choices even when the AI never interacts directly with the target.

Table 4. Connecting the survey activities to the TLPT sequence

Organising phase	Relevant survey activities	Connection developed in this report
Initiation	Planning and scoping	Purpose, scope, intended AI role and client agreement.
Threat intelligence	Reconnaissance and threat-informed preparation	Collection, processing, contextual analysis and scenario relevance.
Penetration testing	Execution; elements of post-execution analysis	Technical activity, practitioner authority and alignment with the agreed scenario.
Closure and reporting	Post-execution analysis; reporting and assessment	Validation, evidence, interpretation and organisational learning. Also, replay, purple teaming and remediation.

A decision chain needs an evidence chain

An AI-generated intelligence summary may influence the scenario selected for testing. That scenario may guide a technical action, which in turn contributes to the final report. At each handover, relevance, authority or evidence can change the meaning of the outcome. A credible engagement therefore needs to preserve the rationale for material AI-supported decisions as they move through the test.

For providers, this means clear responsibilities across intelligence specialists, testers and reviewers. The final conclusion should be traceable back to the relevant threat, the authorised scenario, the activity actually performed and the evidence observed - not simply to the output of the tool that assisted at each stage.

5. More analytical capacity raises the bar for threat fidelity

The survey shows AI adding value through greater analytical capacity: faster threat-intelligence analysis, broader information processing, more effective mapping of tactics, techniques and procedures (TTPs), and support for tooling, scripts and scenarios. The tension is that more analytical capacity also creates more possible answers. The quality question becomes whether those answers remain relevant to the threat the engagement set out to test.

Reported usefulness needs separate evidence of outcomes

Respondents rated the AI tooling included in this analysis at an average of 7.2 out of 10, with a median of 7 and most ratings between 6 and 8. That indicates useful operational value for the use cases respondents selected. Evidence of improved test quality or resilience outcomes would require separate evaluation.

Among respondents using AI for threat actor emulation, 86% (18 of 21) report improved accuracy and realism. The finding points to practical value and also sharpens the need to distinguish technical plausibility from threat fidelity. A convincing or elaborate technique contributes to the test when it remains relevant to the actor and the assurance objective.

Table 5. Reported benefits and limitations in threat actor emulation

Benefits described by respondents	Limitations described by respondents
Faster or broader threat-intelligence analysis	Generic or insufficiently threat-specific behaviour
More effective identification and mapping of TTPs	Reduced contextual interpretation or judgement
Support for realistic tooling, scripts and scenarios	Tooling that is still developing in maturity
Support for adaptive or dynamic attack paths	Difficulty maintaining the agreed threat scenario

AI can produce more answers; practitioners still have to choose the right one

AI can increase the volume of intelligence collected, processed and correlated during preparation. Roundtable participants also highlighted the risk that model and data biases could make different threat actors appear more alike or steer analysis towards generic patterns. Greater scale therefore increases the need to preserve source context and analytical challenge.

Validation involves more than checking whether an output is technically plausible. Practitioners need to ask whether the information is relevant to the organisation, whether proposed TTPs are credible for the actor and whether the scenario still reflects the threat being examined. Intelligence expertise is what turns a larger set of possible connections into a defensible testing plan.

The relationship between threat intelligence and penetration testing becomes more important as AI creates additional routes and techniques. Testers need enough visibility of the original intelligence rationale to judge whether a proposed change remains within the agreed scenario or requires further intelligence assessment or client agreement.

Alternative attack paths need a threat rationale

AI can surface alternative attack paths that expose a useful weakness or reflect an emerging possibility. Technical viability is only part of the decision. A new attack path also needs a clear relationship to the agreed actor and scenario. Roundtable discussions distinguished actor-specific emulation from broader simulation and highlighted the need to explain when a proposed route changes the basis of the test.

Providers should therefore be able to record why an alternative was considered, what evidence supports it and whether the authorised scenario changed. That preserves flexibility without allowing technical novelty to drift away from the threat the engagement is meant to represent.

AI-enabled impersonation raises the ethical threshold

AI can make social engineering far more convincing through cloned voices, synthetic video and highly personalised content. The roundtables identified this as a clear ethical boundary for AI-enabled TLPT. Participants described growing client interest in deepfake techniques, sometimes to demonstrate risk to senior leaders or boards, alongside concern about the intrusion and distress caused when a test uses an identifiable person’s voice, likeness or personal information.

One practical example was to test whether verification or escalation controls would withstand a convincing impersonation by asking an authorised individual to place a controlled call through a red-team channel. This can exercise the relevant control while reducing the personal impact of the test, provided the assurance conclusion is clear about which elements of a full deepfake scenario were exercised.

The wider principle is necessity and proportionality. Highly personalised AI content, voice cloning and deepfakes should have a clear relationship to the threat being emulated and the assurance outcome being sought, with justification, authorisation and safeguards reflecting the additional intrusion or potential harm.

6. How AI authority changes the control model

The authority granted to AI shapes the consequences of its activity and the controls needed around it. The survey separates three distinct questions: what respondents currently permit, what level of autonomy they consider appropriate and the conditions under which they could envisage greater autonomy in future.

Human control sits at different points in the workflow

During execution, practitioner approval before action is the most frequently reported arrangement in the commercial group, while advisory use is the most frequently reported arrangement in the regulated group. Both retain a human decision point, but at different places between analysis and action.

Table 6. Reported authority during execution

Reported arrangement	Commercial	Regulated
AI not permitted	1 (7%)	1 (9%)
AI not currently used	1 (7%)	2 (18%)
AI provides advice only	3 (21%)	6 (55%)
Human approval required before action	8 (57%)	2 (18%)
AI may perform autonomously	1 (7%)	0 (0%)

One commercial response permits autonomous execution; none of the 11 regulated responses does. The important assurance question is therefore where authority transfers: who reviews a proposed action, what information informs that decision, what has already been delegated and how quickly a practitioner can intervene if behaviour departs from expectation.

Bounded autonomy is where most respondents draw the line

Table 7. Highest level of autonomy considered appropriate at the time of the survey

Respondent view	Responses
Assistance to human testers, with independent testing actions excluded	11 (44%)
Limited autonomous actions within tightly defined boundaries	10 (40%)
Substantial autonomous activity with human oversight and intervention	3 (12%)
Fully autonomous TLPT in some circumstances	1 (4%)

n=25

Looking further ahead, 14 of 26 respondents favour human oversight in all cases, 9 envisage limited or tightly controlled autonomous scenarios, and 3 envisage potentially full engagements. Conditions raised include reliability, technical scope controls, monitoring, logging and explicit client or regulator approval. The direction of travel is therefore contested, but the appetite for unconstrained autonomy is limited.

Roundtable participants reinforced that caution. They raised concerns about routine live deployment of highly autonomous or swarm-based testing, while also describing agentic testing outside the regulated TLPT setting. Autonomy, environment and task all matter; the same level of independence can present very different consequences in research, analysis and live execution.

Scope, data and reliability are the leading risks

Table 8. Relative priorities among six AI-related delivery risks

Risk	Ranked in top three	Average rank
Scope and operational control	17/25 (68%)	2.7
Client data and confidentiality	17/25 (68%)	2.8
Accuracy and reliability	16/25 (64%)	3.0
Human oversight and accountability	10/25 (40%)	3.7
AI-generated tooling and code	11/25 (44%)	3.9
Traceability and assurance	4/25 (16%)	4.9

n=25. Respondents ranked all six risks; a lower average rank indicates greater concern.

Respondents rank scope and operational control, client data and confidentiality, and accuracy and reliability as the leading concerns. That pattern supports a consequence-based approach: the stronger the potential effect on systems, sensitive information or the agreed conduct of the engagement, the stronger the case for explicit boundaries, review and evidence.

Control has to exist outside the model

Model safeguards are only one layer of control. Network restrictions, technical scope boundaries and permissions around the workflow can constrain what an AI-enabled system is able to reach or do even when the model itself behaves unexpectedly. For higher-consequence uses, the surrounding architecture may be as important as the behaviour of the model.

Lower autonomy does not necessarily mean lower risk. An advisory workflow can still process sensitive client information through an external model, while a tightly bounded agent can still depend on a third-party reasoning service. Access, data exposure and dependencies therefore need to be assessed alongside autonomy rather than inferred from it.

The client's own AI deployments can also form part of the target or scenario. Prompt injection, weak identity and permissions design, unexpected agent behaviour and resource limits may all shape test design. In those cases, scope needs to distinguish clearly between the AI capability being tested and the AI capability used by the testing team.

As operational authority increases, so does the importance of monitoring, stop authority, clean-up and reconstructable evidence. A reviewer should be able to distinguish autonomous activity from human direction, identify material decisions and understand when and why intervention occurred.

This makes 'human in the loop' too imprecise as an assurance statement. Approval can sit at individual actions, defined stages or material decisions within a bounded setting. Meaningful oversight depends on where the practitioner has authority, what they can see and whether they can stop the activity in time.

Controlled environments and phased approaches can make higher-risk testing safer, but they also raise a representativeness question. Their assurance value depends on how closely they reproduce the identities, dependencies, data flows and operational conditions that matter in the live environment.

7. Assuring more than just the model

The survey shows a mixed tooling landscape. 15 of 26 respondents use more than one tooling category, combining commercial products with custom integrations, locally hosted models, specialist products and internally developed capability. 8 report a formal partnership or strategic relationship with an AI tooling provider. The result is not a common AI stack, but a range of delivery architectures.

The same model can create different assurance problems

Two providers can use the same underlying model and present very different assurance profiles depending on how it is wrapped, connected and controlled. Integrations, data flows, permissions, validation and human oversight can matter more to the resulting service than the model name itself. The provider's control environment may therefore hold much of the substantive capability.

For buyers and assurance stakeholders, this changes the due-diligence question. Assessing an AI-enabled TLPT service increasingly means understanding the complete delivery environment: what the model can see, what can act on its output, where practitioner decisions sit, which dependencies exist and what evidence shows those controls operated as intended.

Roundtable discussions highlighted external models used as reasoning engines, local hosting and changing vendor arrangements. Each configuration creates different dependencies and data flows, and those conditions can change during the life of an engagement. Providers therefore need a current view of the systems, interfaces and responsibilities that sit around the model.

Disclosure should focus on role, data and control

Respondents place greatest emphasis on disclosures that explain consequences: external processing of client data, the parts of the lifecycle using AI, human oversight and permitted autonomy. The pattern suggests that clients gain more decision-useful information from understanding what AI changes about the engagement than from receiving a list of model names.

Table 9. Information selected among the three most important disclosures

Disclosure priority	Responses
External or cloud-hosted processing of client data	21 (81%)
Parts of the TLPT lifecycle that will use AI	17 (65%)
Human oversight and approval controls	13 (50%)
Permitted level of AI autonomy	12 (46%)
Known limitations or risks	6 (23%)
Logging or evidence of AI-supported actions and decisions	5 (19%)
AI tools or models being used	1 (4%)

Respondents selected the three pieces of information they considered most important to disclose. Percentages reflect multiple selections and may sum above 100%.

Only 1 respondent selected the specific AI tool or model among their three highest disclosure priorities. The result reflects relative priority: respondents place role, data exposure and control above model identity when deciding what clients most need to understand.

Approval should follow material changes in exposure or authority

Table 10. Uses selected as most clearly requiring explicit client approval

Approval priority or position	Responses
Submitting client data to an external or cloud-hosted AI service	24 (92%)
Permitting autonomous testing actions	16 (62%)
Any AI use during a live TLPT engagement	10 (38%)
Influencing or changing the agreed threat scenario or attack path	8 (31%)
Supporting social engineering activity	7 (27%)
Executing AI-generated code or payloads against client systems	5 (19%)
Explicit approval should not normally be required	5 (19%)

The clearest approval boundary is external processing of client data, selected by 24 of 26 respondents (92%), followed by autonomous testing actions at 62%. By contrast, 38% selected any AI use during a live engagement. The pattern supports a risk-based approach in which explicit approval becomes more important when AI changes data exposure, operational authority or the agreed conduct of the test.

For providers, this means separating general transparency from approval of higher-consequence uses. Engagement documentation should explain the intended AI role, the information involved, the approval points that apply and how material changes - such as a new model, integration or data flow - will be handled during delivery.

Data governance is part of the test design

The strong preference for explicit approval around external processing of client data brings data governance into the engagement from the outset. Roundtable participants extended that discussion to processing and storage locations, retention, training use, subcontractors and authority for access. These questions can arise in research, drafting and analysis as well as in agentic activity.

Procurement and initiation provide an opportunity to establish those expectations before delivery begins. If a provider later changes a model, integration or data flow, the decision process should make clear whether the original approval still applies and what evidence shows that the agreed boundary remained intact.

Professional accountability also needs to remain visible as the technical environment becomes more complex. Contractual allocation, statutory liability and insurance require specialist consideration, but the engagement still needs clear operational ownership, escalation and decision rights that providers and clients can understand.

The practical implication is a wider unit of assurance. Confidence in AI-enabled TLPT depends not only on the model, but on the service around it: integrations, permissions, data, dependencies, practitioner authority and the evidence linking those elements together. The unit of assurance is increasingly the delivery environment, not the model in isolation.

8. As capability becomes easier to access, judgement becomes more valuable

AI makes more analysis, content and technical capability available to practitioners. It shifts where expertise adds value. The critical skill becomes the ability to judge whether an output is technically sound, relevant to the scenario and safe enough to stand behind in a live assurance engagement.

Validation depends on technical depth

24 of 26 respondents rate validation of AI-generated technical output as essential. Secure data handling, understanding AI limitations, human judgement and manual exploitation also rank highly. AI capability builds on these existing technical foundations, particularly validation, secure data handling, judgement and manual exploitation.

Table 11. Practitioner capability ratings for the following 12 months

Capability	Essential	Moderately important	Not important
Validation of AI-generated technical output	24 (92%)	2 (8%)	0 (0%)
Secure handling of data used with AI	22 (85%)	3 (12%)	1 (4%)
Human judgement and decision-making	21 (81%)	5 (19%)	0 (0%)
Understanding AI limitations and failure modes	21 (81%)	5 (19%)	0 (0%)
Manual exploitation	19 (73%)	6 (23%)	1 (4%)
Governing autonomous or agent-based testing	15 (58%)	10 (38%)	1 (4%)
Developing or adapting AI-enabled offensive tooling	14 (54%)	12 (46%)	0 (0%)
Reviewing AI-generated code	12 (46%)	12 (46%)	2 (8%)
Understanding model behaviour and prompt manipulation	10 (38%)	14 (54%)	2 (8%)
Threat intelligence analysis	9 (35%)	14 (54%)	3 (12%)

Threat intelligence analysis received fewer 'essential' ratings than validation or secure data handling, but most respondents still rate it at least moderately important. The roundtables gave it greater emphasis because of its role in threat relevance and scenario interpretation. Capability therefore needs to be considered in relation to the decisions each role is expected to make.

Intelligence specialists need depth in threat relevance and context; testers need depth in technical behaviour and consequence; reviewers need the judgement to assess the evidence behind a finding. Those roles can share foundations in validation and decision-making without requiring identical expertise.

Generated capability can outpace the experience to supervise it

AI can accelerate the creation of niche tools, post-exploitation utilities and reporting automation. That speed creates a supervision problem: sophisticated capability can be generated more quickly than practical experience can be developed. Secure development, peer review and testing therefore become more important before generated capability reaches a live environment.

Training and development need to preserve foundational skills while adding AI-specific capability. Technical depth helps practitioners recognise failure, challenge generated outputs, understand consequences and know when to stop or escalate. Experience with AI then extends the range of tasks to which that expertise can be applied.

If clients can build the tools, what are they buying from the provider?

Independent challenge, complex and novel scenario design, judgement under uncertainty and cross-market insight may therefore become more important differentiators. The external provider's role is not simply to bring a better tool, but to test assumptions, choose relevant attack paths and provide credible independent assurance over the result.

AI can reduce the effort needed to create capability while adding assurance overhead

AI-enabled delivery can reduce the effort required to create some technical capability, but high-assurance use can still require engineering, compute, hosting, monitoring and additional controls. Roundtable participants raised affordability and access as potential constraints, particularly for smaller providers and when emulating well-resourced adversaries.

The commercial question is how the effort saved in creating a tool or workflow compares with the effort required to validate, govern and operate it responsibly. Lower barriers to creation can coexist with higher assurance costs.

A further risk is convergence. If teams increasingly rely on common models or toolchains, superficially diverse services could still develop shared blind spots. Preserving diversity of technique, independent judgement and scrutiny of scenario assumptions may become part of preserving the value of external challenge.

9. Controls can make a test safer while changing what it proves

AI can accelerate research, decision-making and technical activity, but TLPT also relies on safeguards that can slow or constrain the exercise. That creates a tension at the heart of AI-enabled testing: controls can reduce operational risk while also changing the conditions under which resilience is being tested.

Pace is part of the evidence

AI can compress activity that previously unfolded over a longer period. Practitioner approvals, control-group coordination and other safeguards can introduce pauses that would not exist for a real adversary. Interpretation therefore needs to account for both the outcome and whether the timing observed in the test still represents the threat being claimed.

Survey free-text contributions also highlight the value of documenting autonomous activity and the effect of AI on timings. Where AI acceleration, human approval or test governance materially changes the sequence, that context should be visible so the client can distinguish the pace of the threat from the pace of the exercise.

Risk-tiered pre-authorisation and defined stop conditions are possible ways to preserve accountability while allowing relevant time pressure to remain part of the test. The aim is to place human control where it protects the engagement while preserving the behaviour and timing the exercise is intended to examine.

Containment can reduce risk - and realism

Development environments, cyber ranges, digital twins and phased testing can support examination of higher-risk behaviour under controlled conditions. Their value depends on how faithfully they reproduce the identities, dependencies, data flows and operational processes that would shape the same activity in a live environment.

Testing parts of an attack chain in isolation can reduce exposure and can also remove interactions that make the chain meaningful. End-to-end testing may provide stronger evidence about detection, escalation and response. Safer test designs can reduce representativeness, so the assurance conclusion needs to reflect that trade-off.

It should be noted that certain regulated tests require live production testing and restrict de-chaining unless approval is given by the regulator, for example, for ethical reasons.

Table 12. Questions connecting test design to resilience evidence

Dimension	Question for interpreting the engagement
Scenario relevance	How does the tested activity relate to the agreed threat and intended outcome?
Pace and control	How did approval points, pauses and AI assistance affect the timing of activity and response?
Coverage and dependencies	Which systems, relationships and parts of the attack chain were exercised?
Response and recovery	What evidence supports conclusions about detection, decisions, containment, response or recovery?
Learning	Which findings are supported by the observed activity, and where does uncertainty remain?

The assurance conclusion needs to show what was not tested

The roundtables also raised the risk of leaving relevant AI-enabled threats untested. Tighter controls can reduce testing risk while narrowing what the exercise is able to demonstrate. Proportionality therefore works in both directions: the decision needs to consider the residual risk introduced by the test and the uncertainty created by not exercising relevant behaviour.

For buyers and providers, the practical response is to make those boundaries explicit. The engagement should record the exposure being addressed, the safeguards supporting the chosen design and the questions that remain outside scope. That allows the final resilience conclusion to distinguish observed evidence from assumptions or untested dependencies.

Cross-entity and supply-chain scenarios create the same problem at a larger scale. Cloud services, development pipelines and software dependencies can widen the control group and complicate covertness. Clear records of scope, participation and tested relationships are needed so that governance arrangements do not obscure what the exercise actually demonstrated.

AI therefore changes the operating conditions around TLPT as well as the capability within it. Credible resilience conclusions need to explain how pace, containment, authority and coverage shaped the evidence. Safety remains essential, and its effect on representativeness should be explicit in the assurance conclusion.

10. Acceptance and scrutiny are increasing together

Respondents expect AI to affect delivery, procurement and assurance. Greater use of offensive tooling and specialist skills sits alongside expected increases in procurement requirements, disclosure requests, audit evidence and regulatory attention. The survey therefore points to growing acceptance accompanied by greater scrutiny.

Client support comes with greater assurance expectations

Table 13. Providers' accounts of current client attitudes

Attitude attributed to most clients	Responses
Generally supportive where appropriate controls are in place	13 (50%)
Actively supportive and encouraging use	6 (23%)
Generally cautious and prefer limited use	6 (23%)
Actively opposed	1 (4%)

19 of 26 respondents describe clients as supportive, including 13 whose description explicitly links that support to appropriate controls. 17 expect support to increase. These are provider perceptions rather than direct client responses, and they suggest that acceptance and scrutiny can develop together.

Table 14. Expected increases over the following 12 months

Area of expected increase	Responses / base	Percentage
Use of AI-generated offensive tooling	23/26	88%
Demand for testers with AI-specific skills	23/26	88%
Regulatory attention on AI use during TLPT	22/25	88%
AI-specific procurement requirements	22/26	85%
Client requests for disclosure of AI use	21/26	81%
Use of autonomous or agent-based testing	20/25	80%
Need for audit evidence relating to AI actions	20/26	77%

For providers, that creates a commercial as well as an assurance imperative. Internal governance has to become visible to buyers through service descriptions and engagement records that explain what AI is used for, how data is handled, where authority sits and what evidence supports the result. As AI becomes more normal, demonstrable control may become more important rather than less.

Assurance needs to work across uneven market readiness

Those expectations will land in a market with very different levels of organisational and provider readiness. Roundtable participants, particularly from financial services, highlighted variation across institutions, suppliers and supervisory environments. A common assurance objective may therefore need to be achieved through different technical implementations rather than a single prescribed model.

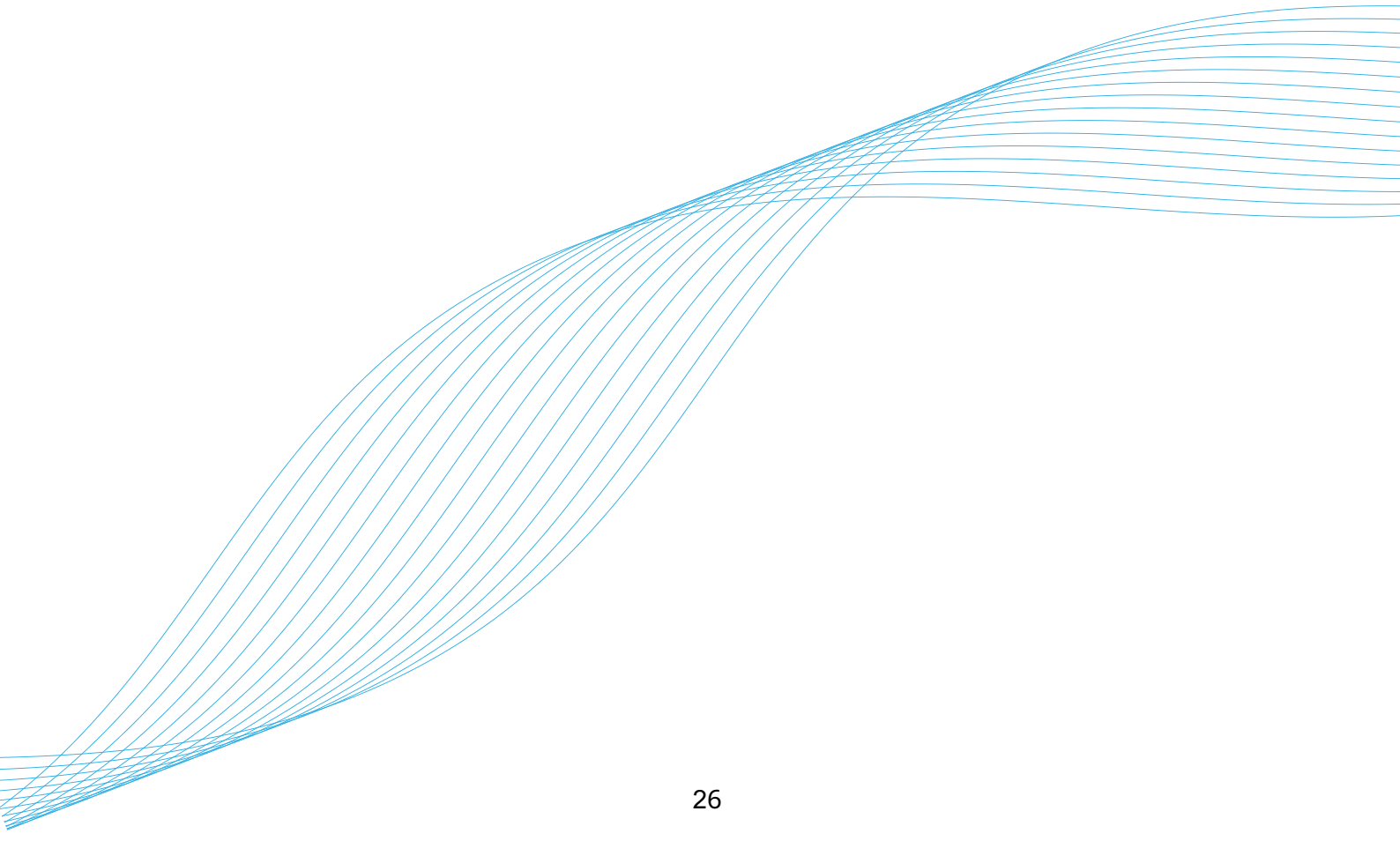
The durable requirement is to define the outcomes that need to be preserved: enough understanding for a buyer to commission and interpret the service, enough competence and control for a provider to deliver it, and enough evidence for the resulting assurance to be credible.

This is especially relevant to supply-chain testing, where large institutions and smaller suppliers may have very different resources and experience. Staged development, shared learning and proportionate expectations can support wider participation, but they should not dilute the core questions of threat relevance, accountability and evidence.

The near-term shift is towards demonstrable control

The survey provides a clear 12-month outlook: respondents expect greater use of AI-generated offensive tooling, stronger demand for AI-specific skills, more procurement and disclosure requirements, more audit evidence and increased regulatory attention. The immediate challenge is to make growing AI use explainable, controlled and assessable within today's engagements.

As AI becomes a routine part of delivery, buyers are likely to care less about the novelty of using it and more about the quality of the service built around it. Outcome-based expectations around role, data, authority, validation and evidence provide a more durable basis for assurance than model names or broad claims of AI maturity.



11.Conclusion: AI is changing how TLPT is delivered and assured

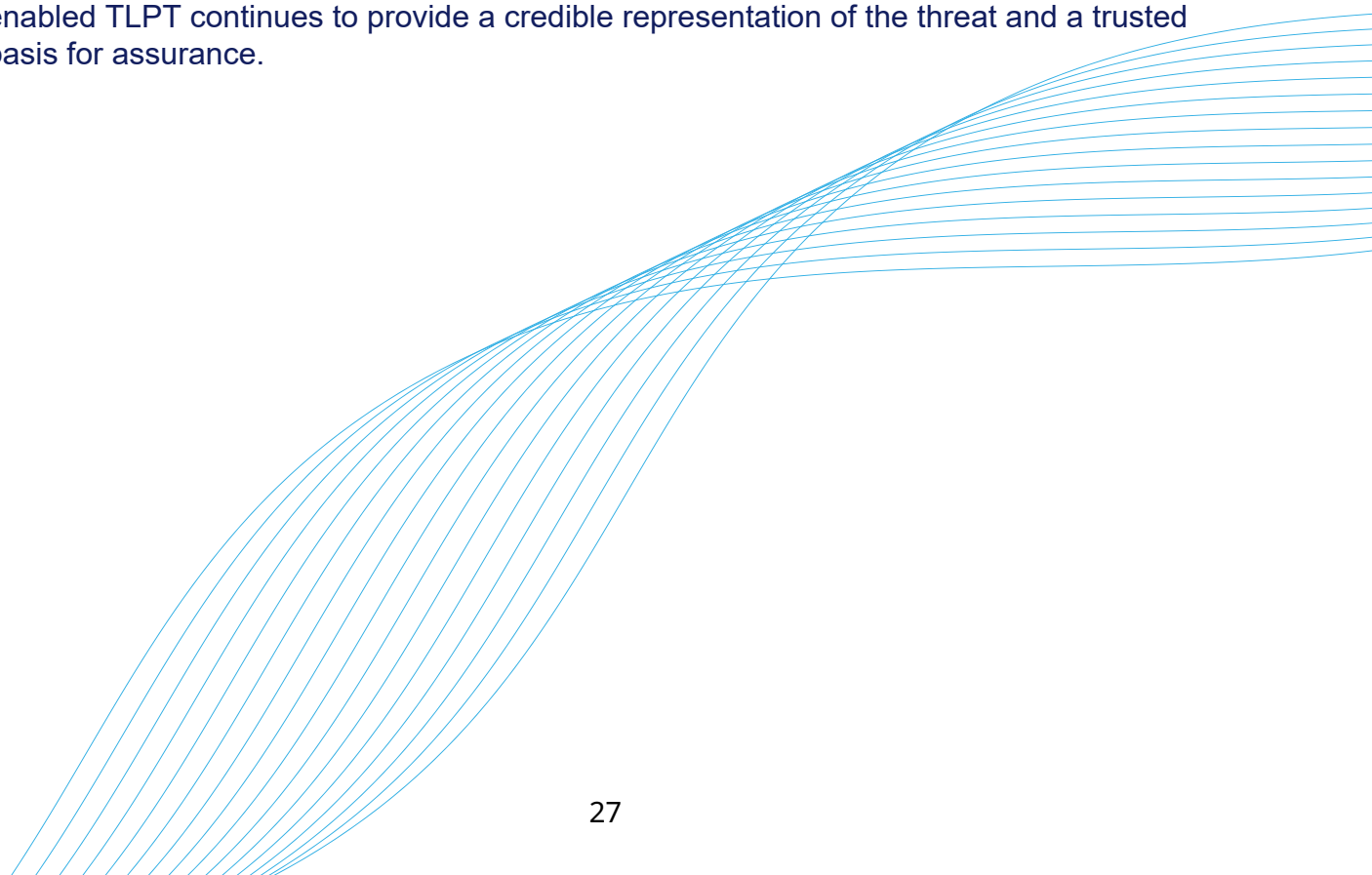
The research points to a broader change in TLPT: AI is beginning to influence where decisions are made, how quickly capability can be created, how threat scenarios evolve, what providers can offer and what evidence is needed to support a credible conclusion. This moves the assurance question closer to the methodology and delivery model itself.

For providers, the challenge is to define AI's role at engagement level: the activities it supports, the information it can access, the authority it receives, the points at which practitioners intervene and the evidence retained. Models, integrations, permissions, data flows and controls all contribute to the delivery environment that needs to be understood and assured.

For intelligence specialists and testers, deeper AI use raises the value of threat context, manual technical skills, validation and supervision. Greater capability creates more options, and professional judgement determines which are relevant, proportionate and safe enough to use. The credibility of the test depends on a defensible link between the threat, the authorised scenario and the activity performed.

For buyers and assurance stakeholders, the priority is decision-useful transparency. A credible service should make it possible to understand what AI changed about the engagement, what data and authority were involved, where human decisions sat, what controls constrained the activity and what evidence supports the resulting conclusion.

General AI governance provides an important foundation around approved use, data protection and human oversight. TLPT adds service-specific questions around threat fidelity, scenario change, technical authority, intervention, test pace and the evidence supporting resilience conclusions. Addressing these questions helps ensure that AI-enabled TLPT continues to provide a credible representation of the threat and a trusted basis for assurance.





CREST.org

© 2026 CREST (International), All Rights Reserved