



Commerce & Contract
Management Institute

How to Buy and Sell AI

A Practical Guide for
Commercial and Contract
Managers

Co-founded by:



Introduction

Artificial intelligence is a rapidly evolving field, with technologies, regulation, good practice and organisational approaches continuing to develop at pace. This guide is therefore intended to be a dynamic and practical resource, rather than a formal or definitive set of instructions.

The information reflects the position and understanding at the time of publication, but readers should be aware that guidance, regulation and available technologies may change.

We welcome and encourage feedback from users. Your comments and experiences will help us keep this guide relevant, useful and up to date as the AI landscape continues to evolve.

About this guide

This guide helps public sector, non-government and commercial organizations navigate the distinct challenges of buying and selling artificial intelligence. It translates a simple framework, the 5Rs (Responsibility, Reliability, Resilience, Regulation and Reasoning), into practical contract requirements for both buyers and suppliers.

It is written for commercial and contract practitioners, not engineers. You do not need a technical background to use it. Where deeper technical material is genuinely useful, such as benchmarks, evaluation metrics and infrastructure detail, it has been moved to the appendices, so it is available to those who want it without getting in the way of those who do not.

Use it in a risk-based way. Not every requirement here belongs in every contract. The controls you need depend on what the AI does and what happens when it gets things wrong. Part I sets out a risk-tier approach, and the checklists in Part II are calibrated to it. Treat the guide as a menu matched to risk, not a mandatory standard.

A living document

AI practice and regulation are moving quickly, so this is a revised edition rather than a final one. The Institute intends to review it periodically and welcomes feedback from practitioners who put it to use. Where this guide states dates or specific rules, confirm the current position for your jurisdiction before relying on them.

Who this guide is for

- Public sector contracting officers and procurement teams acquiring AI
- Commercial buyers evaluating AI suppliers and products
- Proposal and business development teams selling AI
- Legal and contract professionals on either side of an AI transaction

What you'll find

- **Part I** — the 5Rs framework, the AI concepts that change how you contract, including agentic AI and human oversight, and a risk-tier model for matching controls to consequences.
- **Part II** — a bilateral obligations checklist for suppliers, buyers and the parties jointly, with brief context on why each item matters.
- **Part III** — the main negotiation areas: warranties, liability, intellectual property and commercial models, and audit and assurance.
- **Appendices** — a technical reference (A), a risk-tier control matrix (B), and an illustrative AI contract schedule you can adapt (C).

What we mean by “AI”

“AI” is used loosely in the market, and the label often drives price and expectation more than the technology does. For contracting purposes the useful distinction is not marketing category but behavior.

	Genuine AI	Rules-based automation
How it behaves	Learns patterns from data; outputs are probabilistic and can vary for the same input; behavior can change as data or the model changes.	Executes hand-written rules, scripts or macros; outputs are deterministic and repeatable; behavior changes only when someone rewrites the rules.
What can go wrong	Hallucination, bias, drift, opaque reasoning, emergent behavior.	Logic errors and unhandled edge cases, but predictable and testable.
What the contract needs	Performance ranges, bias and drift controls, explainability, human oversight, incident handling.	Logic validation, version control, integration testing, change management.

Some products sold as “AI” are really automation, and some blend both. This matters commercially: paying an AI premium, and accepting AI-grade uncertainty, for what is in fact deterministic automation is a poor bargain, while contracting for automation when the system actually learns leaves real risks uncovered. Establish which you are buying, and contract accordingly.

A note on terminology

For consistency this guide uses Buyer and Supplier throughout. Where an obligation is specific to government or the wider public sector it says Public Sector Buyer. The parties means buyer and supplier jointly. Substitute your own contract's defined terms as needed.

Why these five dimensions matter

AI systems fail in ways traditional software does not. The 5Rs give you a structured way to decide what a given contract actually needs. The first four ask who is accountable, whether the system performs, whether it lasts, and what rules apply. The fifth, Reasoning, cuts across the others: can the system be understood and explained well enough for the decisions it supports.

Responsibility: who is accountable when AI fails?

Traditional software failures are relatively contained. AI introduces new accountability challenges: emergent behavior the vendor did not explicitly program; distributed causation, where a failure may stem from training data, model design, deployment context or user interaction; and new harm types, including discrimination, privacy violations and safety risks, not just technical faults.

Contract implication. Standard warranty and indemnity clauses often do not account for algorithmic harm. Responsibility provisions should address who bears risk when AI produces discriminatory outcomes, makes consequential errors, or operates outside agreed bounds.

When this matters most

- High-stakes decisions (hiring, lending, law enforcement, healthcare)
- Public-facing AI affecting individuals' rights or opportunities
- Regulated activities with compliance obligations
- Where errors could cause physical, financial or reputational harm

Reliability: will this AI perform consistently?

Reliability varies by system type, and each type has its own failure modes. Generative systems produce variable outputs and can state falsehoods convincingly. Predictive systems can degrade quietly as conditions change. Perception systems can perform unevenly across different groups and be fooled by adversarial inputs. What “good enough” means, and how it is measured, has to be defined for the specific use case rather than assumed.

Contract implication. Set performance expectations as measurable acceptance criteria appropriate to the use case and its risk tier, with agreed methods for detecting and remedying degradation. Part II sets these out; the underlying metrics sit in Appendix A.

Resilience: can this AI adapt and recover?

AI degrades over time in ways traditional software does not. Model drift erodes performance as real-world data diverges from training data. Adversaries can craft inputs to fool the system. Models may need retraining to hold performance, and switching suppliers is often harder than for conventional software.

Contract implication. Address ongoing monitoring, the approach to model change and retraining, fallback procedures when AI fails, and exit and continuity if the solution stops meeting your needs.

When this matters most

- Long-term, multi-year commitments
- Rapidly evolving operational environments
- High supplier lock-in risk
- Business-critical AI where degraded performance has serious consequences
- Adversarial environments (fraud detection, cybersecurity)

Regulation: what compliance obligations apply?

The AI regulatory landscape is still settling, and it differs by jurisdiction, so contracts need to handle both current obligations and change over time. As at August 2026 the position includes:

- European Union. Under the EU AI Act, bans on prohibited practices and an AI-literacy duty have applied since February 2025, and obligations on general-purpose AI models (technical documentation, a training-data summary, copyright compliance) since August 2025, with the AI Office's enforcement powers live from August 2026. Article 50 transparency duties, such as disclosing that users are interacting with AI and marking AI-generated content, also apply from August 2026. Most obligations on high-risk (Annex III) systems were deferred by the Digital Omnibus package to December 2027, and product-embedded systems to August 2028; treat these as planning anchors and confirm final adoption.
- United States. The landscape remains fragmented across federal executive action, agency and procurement guidance, and a growing set of state laws. Public sector buyers should work from their agency's current acquisition guidance rather than any single national rule.
- Sector rules. Healthcare, financial services, defense and others impose their own AI-relevant requirements that sit on top of general AI law.

Contract implication. Address current obligations, allocate responsibility for regulatory change, and build in a mechanism to update the contract as rules evolve. Write to the principle (transparency, human oversight, risk management, record-keeping) and reference specific standards as implementation detail, so the contract ages better. Demonstrable compliance can also be a competitive differentiator.

When this matters most

- Public procurement subject to specific AI acquisition policies
- Highly regulated industries (healthcare, financial services, defense)
- Cross-border operations spanning different jurisdictions
- High public visibility, or where non-compliance carries severe penalties

Reasoning and Transparency: can we understand and explain it?

Reasoning is the fifth R, and it runs across the other four. It asks whether a system can be understood and explained to the degree the decision demands: not how the model works internally, but whether the buyer can see what it was trained on, what its known limitations are, why it produced a given output, and how that output should be interpreted. A system nobody can explain cannot be held responsible, its reliability cannot be judged, its drift cannot be diagnosed, and it cannot demonstrate regulatory compliance. Explainability should scale with stakes: a low-risk drafting aid needs little, while a system informing a decision about a person's rights, money or safety needs enough to justify and, if necessary, appeal that decision.

Contract implication. Require transparency proportionate to the use case: disclosure of model type, training-data characteristics and known limitations; explanation of individual outputs where decisions affect people; and logging sufficient to reconstruct what happened after the fact. Treat it as a cross-cutting requirement, not a single clause.

AI types and their risk profiles

The type of AI shapes which of the 5Rs to prioritize and how to structure requirements.

AI type	What to watch for	Contract emphasis
Generative	Variable outputs; plausible but false content (hallucination); hard to test exhaustively.	Output-quality standards, hallucination and toxicity controls, IP and privacy protection, appropriate audit rights.
Predictive	Quiet degradation; fairness and compliance risk in scoring, forecasting and recommendations.	Accuracy expectations for the use case, explainability, fairness testing, regulatory compliance.
Deterministic / rules-based	More predictable, but edge-case and subtle logic failures.	Logic validation, version control, integration testing, change management.
Perception	Uneven performance across groups; adversarial vulnerability; drift in changing conditions.	Performance equity across groups, adversarial robustness, continuous accuracy monitoring.

Maturity matters too. Organizations beginning with AI should focus first on Responsibility and Regulation, setting guardrails before scaling. Those piloting should emphasize Reliability, proving the technology works before expanding. Those in operation should add Resilience for the long term. Where AI becomes mission-critical, all five Rs need sustained attention.

Matching controls to risk: a tiered approach

The most common criticism of AI contract checklists, including earlier versions of this one, is that they read as absolute requirements. In practice a low-risk drafting assistant and a system that helps decide who gets a loan should not carry the same obligations. Tier the use case first, then apply controls in proportion. Place a use case in the higher tier if any single factor points there.

Factor	Lower risk	Higher risk
Consequence of error	Minor, easily reversed, human catches it	Serious, hard to reverse, affects rights, safety or money
Autonomy	Advisory; a human decides and acts	Acts on its own, or humans rubber-stamp in practice
Data sensitivity	Public or low-sensitivity data	Personal, special-category, classified or commercially critical data
Reach and visibility	Internal, small scale	Public-facing, large scale, or politically sensitive

This gives three working tiers. The controls below are indicative starting points; Appendix B maps them across all five Rs.

Risk tier	Typical use cases	Baseline controls
Low	Drafting aids, summarization, internal search, low-stakes automation.	Basic performance expectations, acceptable-use terms, supplier self-attestation, standard data protection, light human review.
Medium	Customer-facing assistants, operational forecasting, triage and prioritization.	Measurable acceptance criteria, bias and drift monitoring, explainability on request, defined human review, third-party assurance, incident reporting.
High	Decisions affecting rights, safety, livelihood or money; agentic systems that act; regulated high-risk uses.	Rigorous acceptance and fairness testing, mandatory human decision and override, full logging and explanation, independent audit, strict incident timescales, elevated liability and insurance, kill-switch and continuity.

Agentic AI: when the system acts, not just advises

Most contracting practice assumes AI recommends and a person acts. Agentic AI breaks that assumption: it plans, uses tools, and takes actions in the world, sometimes across multiple cooperating agents, with limited step-by-step human involvement. An agent that can send messages, move money, change records or trigger downstream systems is exercising delegated authority, and the contract has to treat it that way.

Agentic deployments need controls that advisory systems do not:

Control	Why it matters
Approval thresholds: Actions above defined limits (value, sensitivity, irreversibility) require human authorization before execution	Keeps consequential or irreversible actions under human control while letting routine actions flow.
Kill switch and halt: A reliable, tested means to suspend the agent immediately, with defined authority to use it	When an agent misbehaves, you need to stop it in seconds, not raise a support ticket.
Scope and tool constraints: Explicit limits on the systems, data and actions the agent may access or perform	Bounds the blast radius; an agent should not be able to act outside its intended remit.
Action logging: A complete, tamper-evident record of what the agent did, when, and on what basis	Accountability and incident analysis are impossible without a reconstructable trail of autonomous actions.
Multi-agent accountability: Clear allocation of responsibility where several agents or vendors interact	Prevents gaps where no single party owns a failure produced by interacting agents.

Human oversight: keeping people in control

Human oversight is the counterweight to autonomy, and it needs to be designed rather than assumed. Oversight can be in the loop (a human decides before the system acts), on the loop (a human monitors and can intervene), or out of the loop (the system acts alone). The higher the risk tier, the more the design should sit toward in-the-loop for consequential decisions.

Requirement	Why it matters
Review before consequential decisions: For higher-risk uses, a competent human reviews and decides before the AI's output takes effect on a person	Preserves genuine accountability and meets emerging legal duties for human oversight of high-risk AI.
Genuine override: Humans can reject or change AI output, and the workflow makes that realistic rather than nominal	Oversight that cannot practically say no is not oversight; automation bias makes real override essential.
Appeal and redress: People affected by an AI-influenced decision can question it and seek human re-review	Fairness, and often law, require a route to challenge decisions that affect individuals.
Competent reviewers: Reviewers understand the system's limits well enough to judge its output, not just pass it on	Rubber-stamping by under-informed reviewers is a common and dangerous failure of oversight.

A foundation: data and digital literacy

Before procuring, make sure your organization can actually evaluate and oversee what it buys. The questions are simple but decisive: can your team judge supplier claims about accuracy, bias and performance; do requirement-writers know what “good” looks like for this use case; can users recognize when the AI is wrong; and can your people monitor performance and spot degradation after go-live.

- Invest in AI training for procurement, technical and end-user teams
- Bring in expertise where you lack it, through advisors or hiring
- Do not procure AI you cannot evaluate or oversee

Why this matters

Without adequate literacy you cannot write good requirements, evaluate proposals, negotiate sound terms or manage performance after award. The supplier's information advantage will otherwise leave you exposed.

Part II: Bilateral Obligations Checklist

I. Supplier obligations

A. Performance and reliability

Addresses Reliability, Resilience and Reasoning. Performance guarantees must account for AI's inherent variability while remaining enforceable. The aim is objective, use-case-appropriate criteria rather than universal numbers.

Requirement	Why it matters
Acceptance criteria: Performance thresholds appropriate to the use case and risk tier, agreed as measurable acceptance tests	Without agreed, measurable baselines, disputes about “acceptable performance” become subjective. Set the level by use case, not a blanket figure.
Fit-for-purpose evaluation: Evaluation against tasks representative of the actual use, using industry-standard benchmarks appropriate to the model type and use case as reference only	General capability leaderboards do not prove fitness for your task; representative testing does. See Appendix A for the underlying metrics.
Degradation remedies: Alerts when performance drops below threshold, with remediation inside an agreed timeframe	Both parties know immediately when performance slips and what response to expect, preventing prolonged substandard service.
Drift detection: Ongoing monitoring for model drift with predefined intervention points	Drift is often invisible until it causes real problems; early detection allows intervention before failures occur.
Responsiveness: Response-time expectations suited to the use (real-time versus batch)	Misaligned latency expectations can render capable AI operationally useless; set them to the use case.
Accessibility: Compliance with current accessibility standards, with a mechanism to adopt successors as they emerge	The principle is equitable access; the specific standard is implementation detail that will change over time.

B. Model versioning and change

Addresses Resilience. AI changes frequently, and changes can silently alter behavior. Consolidate versioning, rollback and update testing so the same ground is not covered twice elsewhere.

Requirement	Why it matters
Version control and audit trail: A complete record of model versions, what changed, why, and the performance impact.	AI changes often lack code-level trails; documentation enables root-cause analysis when problems occur.
Update testing: Validation of updates against agreed criteria before full deployment	Catches issues that test environments miss before they reach production users.
Rollback: The ability to revert to a previous stable version, with clear criteria and authority to trigger it	When an update degrades performance, fast, pre-agreed reversion limits user impact and prevents disputes mid-incident.
Compatibility: Interface stability so model changes do not break dependent systems	Model changes can alter inputs and outputs; maintaining compatibility prevents integration failures.

C. Data quality and provenance

Addresses Reliability, Regulation and Reasoning. Buyers cannot assess AI risk without understanding the data behind it.

Requirement	Why it matters
Training-data documentation: Datasheets covering source, collection method, preprocessing and known limitations	Buyers cannot evaluate whether AI will work in their context without understanding its training data.
Provenance and lineage: Documented provenance for training and validation data, including licensing basis	Provenance supports root-cause analysis and guards against IP and privacy contamination in the training set.
Bias documentation: Analysis of known biases in the data, with mitigation	All data carries bias; documenting it turns an invisible risk into a managed one the buyer can weigh.
Refresh: Data refresh appropriate to the domain, with impact assessment	Models trained on stale data degrade; buyers need clarity on refresh frequency, effect and cost.

D. Security and data protection

Addresses Responsibility and Regulation. AI systems process sensitive data at scale and face novel attacks, including prompt injection, model inversion and adversarial inputs.

Requirement	Why it matters
AI-specific threat controls: Protection against prompt injection, model inversion, data extraction and adversarial inputs	These attacks do not exist for conventional software; standard security controls do not fully cover them.
Access and segregation: Least-privilege access, and segregation of buyer data from other tenants and from training use	Prevents unauthorized access and stops buyer data leaking into shared models or other customers.
Encryption and handling: Encryption in transit and at rest, with data-handling matched to classification	Baseline protection for sensitive and, for public sector, classified information.
Supply-chain security: Assurance over the security of upstream models, components and infrastructure	AI systems inherit the vulnerabilities of everything they are built on; see third-party dependencies below.

E. Privacy and personal data

Addresses Regulation and Responsibility. Privacy deserves first-class treatment: AI both consumes personal data and can expose or reconstruct it. Terms here should reflect the buyer's data-protection regime (for example GDPR or equivalent).

Requirement	Why it matters
Lawful basis and purpose limitation: Personal data processed only on a lawful basis and for specified purposes	Using personal data outside its lawful basis exposes the buyer to enforcement and reputational harm.
Data minimization: Only the personal data necessary for the purpose is collected and processed	Minimization reduces breach impact and is a core legal requirement in most regimes.
No training on personal or buyer data without consent: Supplier will not use buyer or personal data to train or improve models without explicit, documented	Prevents the buyer's confidential and personal data being absorbed into models serving others.

Requirement	Why it matters
Right to erasure and rectification: Mechanisms to delete or correct personal data, including its effect on models where relevant	Individuals' deletion and correction rights must be honored, which is non-trivial once data has influenced a model.
Impact assessment: Support for data-protection impact assessments for higher-risk processing	Individuals' deletion and correction rights must be honored, which is non-trivial once data has influenced a model.

F. Fairness and bias

Addresses Responsibility and Regulation.

Requirement	Why it matters
Bias testing: Fairness testing appropriate to the use case, including performance across affected groups, on a defined cycle	Aggregate accuracy can hide serious disparities for particular groups; tested fairness makes them visible.
Mitigation and thresholds: Agreed action where disparities exceed acceptable bounds	Testing without a duty to act on the results changes nothing.

G. Reasoning, explainability and transparency

Addresses Reasoning and Responsibility. Scale the depth of explanation to the stakes of the decision.

Requirement	Why it matters
Disclosure: Disclosure of model type, key training-data characteristics and known limitations	Buyers cannot manage risk they cannot see; basic transparency is the precondition for every other control.
Output explanation: Explanations of individual outputs where decisions affect people, at a level a caseworker and an affected person can use	People have a growing right to understand decisions about them, and reviewers cannot oversee what they cannot interpret.

Requirement	Why it matters
Decision logging: Logging sufficient to reconstruct what the system did and why	After-the-fact accountability, incident analysis and audit all depend on a durable record.

H. Third-party and foundation-model dependencies

Addresses Resilience, Regulation and Responsibility. Most AI products are built on someone else's model or platform. Those dependencies carry risk that must flow through the contract.

Requirement	Why it matters
Sub-processor and sub-model disclosure: Disclosure of the foundation models, sub-processors and key third parties in the stack	You cannot assess or govern dependencies you do not know exist.
Notice of upstream change: Advance notice of material upstream model changes, deprecation or withdrawal	An upstream change can alter behavior or end availability overnight; notice lets you test and plan.
Continuity: Defined options if an upstream model is withdrawn or materially changed	Protects the buyer from being stranded by a decision made outside the contract.
Flow-down: Relevant security, privacy and compliance obligations pass through to sub-processors	A dependency that is not bound by the same obligations is a gap in the buyer's protection.

I. Agentic controls (where the system acts autonomously)

Addresses Responsibility and Resilience. Where the deployment is agentic, apply the controls from Part I, principally approval thresholds for consequential actions, a tested kill switch, scope and tool limits, and complete action logging. Include them as contract obligations, not design aspirations.

II. Buyer obligations

AI success is bilateral. Many failures trace to buyer-side gaps: poor data, unclear scope, absent oversight or untrained users. These obligations are the buyer's side of the bargain.

Requirement	Why it matters
Data provision and governance: Provide data of adequate quality, with clear rights to use it, and govern it over the contract life	AI can only perform on the data it is given; poor or improperly licensed data undermines the whole system.
Clear scope and intended use: Define the intended use, and the uses that are out of bounds	Off-label use is a leading cause of AI harm and disputed liability; scope defines where the system is warranted to work.
Human oversight resourcing: Provide competent reviewers and the time for genuine oversight, per the risk tier	Oversight obligations placed on the supplier mean little if the buyer does not resource the human side.
Literacy and training: Train the people who build requirements, use outputs and monitor performance	Under-prepared users cannot recognize error or manage the system, re-opening the information asymmetry.
Timely cooperation and change: Give notice of changes in use, data or environment that affect performance	AI tuned to one context can fail silently when the context shifts without the supplier's knowledge.

III. Joint obligations

A. AI incident management

Addresses all five Rs. AI incidents are not just outages: they include harmful, biased or unsafe outputs, privacy breaches, security compromises and unexplained behavior. These deserve a dedicated, pre-agreed process rather than being folded into general service management.

Classify incidents by type and severity, and attach reporting and response timescales to severity and risk tier.

Severity	Illustrative examples	Response expectation
Critical	Safety risk, unlawful or seriously harmful output, personal-data breach, agent acting outside authority.	Immediate containment (including halt or rollback), notify the other party without undue delay, and meet any regulatory or data-subject notification duties.
Major	Material performance failure, significant bias affecting a group, security event.	Prompt notification and remediation within an agreed window, with root-cause analysis to follow.
Minor	Localized errors or degradation within tolerance, caught by human review.	Log, monitor and address through routine service management.
Requirement		Why it matters
Classification and triggers: Agreed definitions of what counts as an AI incident and at what severity		Shared definitions prevent the disputes about “whether this even counts” that waste time mid-incident.
Reporting timescales: Notification windows tied to severity and to any legal duties		Late notice removes the buyer's ability to contain harm and can itself breach the law.
Roles and authority: Named responsibilities, including who may halt or roll back		Incidents are made worse by uncertainty over who can act; decide it in advance.
Post-incident review: Structured review and corrective action, with lessons fed back into the system		Prevents recurrence and turns incidents into improvement rather than blame.

B. Ongoing monitoring and model change management

Addresses Reliability and Resilience. This consolidates monitoring, drift response and retraining into one place to avoid the repetition reviewers noted across earlier drafts.

Requirement	Why it matters
Shared monitoring: Agreed metrics, dashboards and review cadence, visible to both parties	Shared visibility catches degradation early and aligns both parties on the same picture of performance.
Change management: A defined process for retraining, updates and model changes, with impact assessment and buyer notice	Model change is the main source of silent behavioral shift; managing it protects reliability and trust.
Continuous improvement: Regular incorporation of operational insight into the system	Keeps a probabilistic system fit for purpose as conditions move.

C. Governance and issue resolution

Addresses Responsibility and Regulation. Consolidates joint governance, escalation and oversight.

Requirement	Why it matters
Joint governance: Regular joint meetings with clear decision authority and shared performance reporting	AI needs active joint management; a governance forum turns problems into decisions rather than disputes.
Escalation and resolution: A defined escalation path with timeframes before formal dispute mechanisms	Early, structured escalation resolves most issues before they become disputes.
Independent oversight where warranted: For higher-risk uses, ethics or oversight review proportionate to impact	High-impact AI benefits from a check beyond the delivery team; scale it to risk, do not impose it everywhere.
Lines of accountability: Clear separation of day-to-day management, risk and compliance, and independent	Distinct lines of accountability stop risk ownership falling through the gaps.

Part III: Key Negotiations

Four areas generate most of the negotiation in AI contracts: how performance is warranted, how liability is shared, who owns and licenses the IP, and how transparency is balanced against commercial protection. In each, the theme is proportion, matching the term to the use case and its risk tier.

A. Performance warranties

Warrant performance in ways that reflect how AI actually behaves, rather than importing deterministic guarantees that no honest supplier can meet.

Requirement	Why it matters
Use-case-appropriate standards: Standards expressed for the specific task and risk tier, using probabilistic language and confidence where relevant, rather than a single headline figure	Realistic, use-case standards are enforceable; blanket numbers invite either refusal or meaningless commitments.
Graduated consequences: Escalating responses across minor, significant and critical degradation	Not every shortfall warrants the same response; graduation distinguishes monitoring from remediation from exit.
Shared-risk and incentives: Joint responsibility where success depends on both parties, with aligned incentives	AI performance depends on buyer data and use as well as supplier quality; shared models align both to succeed.
Improvement triggers: Defined points at which improvement work becomes mandatory	Clear triggers turn “should improve” into an objective obligation.

B. Tiered liability

Requirement	Why it matters
Risk-based classification: Liability scaled to the risk tier of the application	High-risk AI affecting life, liberty or livelihood warrants different liability than low-risk automation.
Impact-based caps: Caps proportional to realistic potential harm and system criticality	Aligns protection with actual exposure, and keeps liability insurable and the deal viable.

Requirement	Why it matters
Shared triggers: Liability shared according to root cause where both parties contributed	Where failure has multiple causes, shared liability is fairer and more defensible than shifting all risk one way.
Insurance: Coverage appropriate to the risk category	Provides financial assurance that obligations can actually be met if a failure occurs.

C. Intellectual property and commercial models

AI raises IP questions conventional software does not: who owns the outputs, whether training data was lawfully sourced, and what happens to buyer data used in the system. It is also a mistake to assume one ownership model fits all.

Requirement	Why it matters
Background and foreground IP: Clear separation of pre-existing supplier IP from developments made for the buyer	Prevents disputes over what the buyer has acquired versus what the supplier retains.
AI output ownership: Explicit terms on who owns model outputs, and acknowledgment that outputs may attract limited or no IP protection in some jurisdictions	Output ownership is genuinely unsettled; silence creates disputes and false assumptions of exclusivity.
Training-data provenance and indemnity: Warranties on lawful sourcing of training data, with indemnity for third-party IP or privacy claims	The buyer inherits risk from how the model was trained; provenance and indemnity move that risk to the party that controls it.
Buyer data protections: Limits on the use of buyer data, including no training without consent, and return or deletion on exit	Keeps the buyer's data and any derived value from being absorbed into the supplier's wider offering.

Choosing a commercial model. The old default, that the buyer owns whatever it pays to develop, is only one option, and often not the most workable on modern platform-based AI. Weigh the range against total cost of ownership, continuity and control:

Model	What it means	Best suited to
Buyer ownership	Buyer owns custom developments and configurations (for public sector, Crown or unlimited rights).	Bespoke builds, sensitive public sector work, cases where independence from the supplier is paramount.
Supplier retention with broad licence	Supplier retains IP; buyer gets broad rights, potentially perpetual use, modification and continuity protections.	AI built on the supplier's platform and shared models, where full transfer is impractical; often the more viable model.
Hybrid / field-of-use	Ownership split by component, or licensed within a defined field of use, with escrow for continuity.	Mixed builds where some elements are bespoke and some are core supplier platform.

Match the model to the deal rather than defaulting to transfer. What protects the buyer is not ownership as such but assured continuity, appropriate use rights, and protection of its own data.

D. Confidentiality, audit and assurance

Transparency has to be balanced against legitimate commercial protection, and audit rights in particular should be proportionate. A demand for “full system access” is usually unrealistic and rarely necessary; assurance can be achieved in less intrusive, and more negotiable, ways. Scale the mechanism to the risk tier.

Assurance level	Mechanism	Typical fit
Lighter	Supplier self-attestation and evidence packs against agreed criteria.	Lower-risk uses, standard commercial products.
Moderate	Independent third-party audit or certification; summary reporting to the buyer.	Medium-risk uses; regulated contexts wanting objective assurance without exposing trade secrets.
Stronger	Defined direct audit or inspection rights, security-cleared review where sensitivity requires it.	High-risk and public sector uses where the stakes justify deeper access.
Requirement		Why it matters
Proportionate audit rights: Audit and inspection scaled to risk, not a blanket right of full access		Full-access demands are often unworkable for suppliers and unnecessary for the buyer's actual assurance need.
Trade-secret protection: Safeguards for genuinely proprietary methods within any audit or transparency regime		Suppliers have legitimate IP to protect; respecting it keeps them at the table and keeps terms realistic.
Public summary reporting: For public sector, transparency to the public through summaries rather than disclosure of confidential detail		Delivers public accountability without forcing disclosure of legitimately confidential information.

E. Contract success factors

Beyond the clauses, a few practices consistently separate AI contracts that work from those that do not.

- **Before contract:** an AI-readiness assessment, realistic expectation-setting, a proof-of-concept with defined success criteria, and stakeholder alignment.
- **During delivery:** active joint governance, open performance reporting, early and collaborative issue management, and continuous learning.
- **For the long term:** planned technology refresh, knowledge retention, a clear exit strategy with data portability, and built-in flexibility for regulatory change.

Successful AI contracting means moving beyond traditional IT procurement to address what makes AI different: probabilistic performance, the potential for bias, drift over time, autonomous action, and fast-moving regulation. The 5Rs give structure to that complexity, but the outcome still depends on partnership between buyer and supplier, each bringing the right capabilities and meeting their obligations.

Use the checklists as starting points, not mandatory requirements. Tier them to your use case, its risk profile and your organization's maturity. What matters most is thoughtful application of the 5Rs, translated into specific, enforceable provisions that protect both parties while enabling the AI to succeed.

Keeping this guide current

Given the pace of change, the Institute will review this guide periodically and update it as practice, technology and regulation move. Practitioners are encouraged to share what worked, what did not, and what is missing, so future editions reflect real use. Check for the current edition before relying on any specific date, rule or standard cited here.

How to use this appendix

This material is reference, not a prescriptive standard. It is here for practitioners who want to understand the technical terms suppliers use, and for technical colleagues supporting an evaluation. Do not lift these metrics into a contract as pass/fail thresholds; agree acceptance criteria for your specific use case instead, with technical advice. Naming a benchmark you cannot interpret can do more harm than good.

Model evaluation benchmarks

Public benchmarks measure general capability, not fitness for your task. They are useful for orientation and comparison, not as acceptance tests. Common examples include general-knowledge and reasoning suites (such as MMLU), grade-school and mathematical reasoning sets (such as GSM8K), commonsense reasoning (such as HellaSwag) and truthfulness measures (such as TruthfulQA). Older statistical measures such as perplexity or BLEU are weak indicators of instruction-tuned model quality. Benchmarks also age and can leak into training data, so treat any single score with caution and prefer evaluation on tasks representative of your actual use.

Other metrics you may encounter

- Accuracy, precision, recall, F1. Classification quality measures; which matters depends on whether false positives or false negatives are more costly in your use case.
- Latency and throughput. Speed of response and volume handled; real-time uses need low latency, batch uses can tolerate more.
- Drift measures. Statistical monitoring of how far current data and performance have moved from the baseline.

Infrastructure and standards

Higher-risk or large-scale deployments may involve requirements on hosting and data residency, scalability and availability, integration and interoperability, and accessibility (for example WCAG at the current version, with a mechanism to adopt successor standards). Treat these as implementation detail to be specified with technical and security colleagues, proportionate to the risk tier.

Appendix B: Risk-Tier Control Matrix

An at-a-glance mapping of the 5Rs to the three risk tiers from Part I. Use it to sanity-check that a contract's controls are proportionate: reach for the High column only where the use case genuinely sits there.

Dimension	Low	Medium	High
Responsibility	Standard liability and acceptable-use terms.	Defined liability by cause; named accountability.	Elevated, impact-based liability; insurance; independent oversight.
Reliability	Basic performance expectations.	Measurable acceptance criteria; drift monitoring.	Rigorous acceptance and fairness testing; strict remediation.
Resilience	Standard support and updates.	Change management; monitoring; rollback.	Continuity guarantees; kill switch; exit and portability.
Regulation	General compliance warranty.	Documented compliance; DPIA where relevant.	Full regulatory evidence; conformity where required; audit.
Reasoning	Basic disclosure of type and limits.	Explanation on request; decision logging.	Full explanation of individual decisions; complete audit trail.
Human oversight	Light human review.	Defined review points; genuine override.	Human decision before effect; override and appeal; competent reviewers.

Appendix C: Illustrative AI Contract Schedule

A starting outline for an AI-specific schedule or annex. It is illustrative, not model wording; adapt it to your contract, jurisdiction and the use case's risk tier, and take legal advice. The point is to make sure each area has been addressed somewhere, not to impose a fixed structure.

Schedule section	What to capture
1. Definitions	AI system, model, outputs, training data, agentic action, incident; align with the main agreement.
2. Use case and risk tier	The intended use, out-of-scope uses, and the assessed risk tier that calibrates the rest.
3. Performance and acceptance	Use-case acceptance criteria, evaluation method, degradation remedies, responsiveness.
4. Data and privacy	Data provision and rights, provenance, minimization, lawful basis, no-training-without-consent, erasure.
5. Security	AI-specific threat controls, access and segregation, encryption, supply-chain assurance.
6. Transparency and reasoning	Disclosure, output explanation, decision logging, proportionate to stakes.
7. Human oversight	Oversight model, review points, override and appeal, reviewer competence.
8. Agentic controls	Approval thresholds, kill switch, scope and tool limits, action logging (where applicable).
9. Third-party dependencies	Sub-processor and sub-model disclosure, upstream-change notice, continuity, flow-down.
10. Incident management	Classification, severity, reporting timescales, roles, post-incident review.
11. Monitoring and change	Shared metrics and cadence, change and retraining process, rollback.
12. IP and commercial model	Background/foreground IP, output ownership, provenance and indemnity, chosen commercial model.
13. Audit and assurance	Assurance level and mechanism proportionate to risk; trade-secret protection.
14. Liability and insurance	Risk-based classification, impact-based caps, shared triggers, insurance.
15. Exit and continuity	Termination, data portability and deletion, transition assistance.

About Commerce & Contract Management Institute

The Institute seeks to improve the world through higher standards in buying and selling. Our rigorous, practical research and insights, both relevant and useful, shape global policy and practice. We help society by driving up standards for the exchange of goods and services, resulting in better trading outcomes in both the private and public sectors. As a not-for-profit organization, we were founded, and are supported, by World Commerce & Contracting and NCMA.

www.ccm.institute | Tim Cummins, Executive Director |
t.cummins@ccm.institute | General or media enquiries:
info@ccm.institute



CCM Institute was co-founded and is supported by:

National Contract Management Association

NCMA is a thriving community of over 100,000 contract management professionals globally. Dedicated to fostering a globally recognized contract management profession and strengthening its connections with related acquisition communities, NCMA serves a diverse membership spanning both the public and private sectors.

www.ncmahq.org

World Commerce & Contracting

WorldCC is a not-for-profit association dedicated to helping its global members achieve high performing and trusted trading relationships. With 80,000 members from over 20,000 companies across 180 countries worldwide, the association welcomes everyone with an interest in better contracting: business leaders, practitioners, experts and newcomers.

www.worldcc.com

