Chapter 18. Using the abacus and periodic table for AI performance metrics and safety metrics

Introduction

Artificial intelligence is becoming more capable very quickly. That creates two different questions. The first is performance: what can the system do? The second is safety: can the risks created by that capability be seen, bounded, controlled and corrected?

These questions are connected, but they should not be mixed together. Performance and safety are different systems. They have different nouns, different failure modes and different rates of change. A system can become much better at reasoning, coding, tool use or autonomous action while the controls around it improve more slowly.

This chapter therefore treats them in order. It looks first at performance and completes that analysis. It then turns to safety and completes that analysis. Only after both are visible do we compare them and ask whether safety is keeping pace with capability.

The comparison is broad. It looks at the US and Chinese frontier-AI ecosystems rather than ranking individual laboratories. The purpose is visibility: to see where capability sits, where safety sits, and where the gap between them may be becoming important.

How This Chapter Unfolds

1. **Performance and safety are different -** Separates what AI can do from the controls required around what it can do.
2. **Performance: the nouns -** Defines the recurring features used to describe AI performance.
3. **Performance: the Periodic Table and current position -** Shows how performance changes by level and where the US and Chinese frontier ecosystems appear to sit.
4. **Safety: the nouns -** Defines the recurring features used to describe safety and risk control.
5. **Safety: the Periodic Table and current position -** Shows how safety changes from local controls to national, international and global control, and where the two ecosystems appear to sit.

6. **The gap between performance and safety -** Compares the completed performance and safety pictures and identifies the important mismatches.
7. **What this means for people working on AI -** Returns to the human brain, institutional architecture and the pilots as part of the practical safety system.
8. **What this means for AI -** Explains why increasing capability must remain inside controls that can observe, constrain and correct it.

1. Performance and Safety Are Different

Performance asks what an AI system can do. Safety asks whether the risks created by what it can do can be controlled. A system can be highly capable and poorly controlled. It can also be heavily constrained while remaining only moderately capable.

The distinction becomes more important as AI moves beyond answers on a screen. Reasoning, coding and multimodal capability now combine with tools, software environments, agents and physical systems. The consequence of an error therefore depends not only on the answer the system gives, but on what it can do next.

For that reason, the chapter does not move back and forth between performance and safety. It completes the performance picture first. It then completes the safety picture. The two are compared only at the end.

2. Performance: The Nouns

Performance needs a stable set of features before levels can be compared. The following nouns describe the main forms of capability used in this chapter.

Performance and Capability Nouns. Exhibit 18.1

Noun	Meaning
Reasoning	The ability to infer, analyse, plan, solve unfamiliar problems and connect multiple dependent steps coherently.
Knowledge	The breadth, depth, retrieval and appropriate application of learned or externally retrieved information.
Coding	The ability to generate, understand, debug and modify computer code.
Software Engineering	The ability to work across larger software systems, repositories, architecture, testing, integration, maintenance and deployment.
Multimodal Capability	The ability to understand and combine text, image, audio, video and other sensor or machine inputs and outputs.
Context / Memory	The ability to preserve, retrieve and use relevant information across longer interactions, tasks, projects and time periods.
Tool Use	The ability to select, operate and coordinate external tools, software, databases, APIs and computational resources.
Agentic Capability	The ability to pursue goals across multiple steps, tools and changing conditions with diminishing immediate human direction.
Speed / Latency	The rate at which useful cognitive work can be performed and returned, including responsiveness under demanding workloads.
Cost / Efficiency	Useful capability delivered per unit of compute, money, energy or human supervision.
Accuracy / Factuality	The ability to produce outputs that correspond reliably to evidence and reality, including calibration when the answer is uncertain.
Instruction Following	The ability to understand the intended task, constraints, permissions and requested form and act accordingly.
Helpfulness / Task Effectiveness	The ability to produce outputs or actions that materially advance the legitimate purpose of the user or system.
Physical / Environmental Capability	The ability to affect the external world through robotics, laboratories, vehicles, infrastructure, devices or other actuators.

Physical / Environmental Capability is included deliberately. AI is beginning to connect cognition to machines, laboratories and infrastructure. A performance table that stops at text and software would miss an important part of where capability may move next.

3. Performance: The Periodic Table and Current Position

The Performance Periodic Table is not a seven-point quality score. The same noun remains in each row or column, while its scope, complexity, autonomy, reach and consequence change by level. Higher levels therefore mean a wider operating scope, not simply a better benchmark result.

Periodic Table of AI Performance and Capability. Exhibit 18.2A

AI Level	Reasoning	Knowledge	Coding	Software Engineering	Multimodal Capability	Context / Memory	Tool Use
VII	Sustain cross-domain, global and intergenerational reasoning about civilisation-shaping systems.	Integrate global knowledge continuously across disciplines, cultures and generations.	Design, build and adapt civilisation-scale digital infrastructure with sustained autonomy.	Continuously engineer and re-engineer global critical digital systems.	Maintain civilisation-scale situational understanding across digital and physical modalities.	Maintain long-duration, globally distributed memory while preserving provenance and correction.	Coordinate global digital and physical tool systems while remaining governable.
VI	Reason across institutions, economies and strategic actors at national scale.	Support institution- or nation-scale knowledge systems and scientific discovery.	Operate across major software estates, infrastructure and cyber-physical systems.	Manage institution-scale software systems and interdependent infrastructure.	Fuse large-scale multimodal feeds across organisations and physical environments.	Maintain institution-scale operational memory with selective retrieval and revision.	Operate tool ecosystems spanning institutions, infrastructure and strategic domains.
V	Model complex systems, agents and second-order consequences.	Maintain broad cross-domain knowledge with live retrieval and synthesis.	Plan and execute large multi-component software changes with tools and agents.	Run extended engineering projects with planning, tools, testing and recovery.	Reason over persistent video, audio, documents, sensors and tools as one system.	Use persistent memory across extended goals, changing conditions and multiple agents.	Discover, sequence and supervise tools autonomously over extended tasks.
IV	Integrate multiple domains, constraints and evidence sources.	Integrate knowledge across multiple domains and sources.	Work across repositories, tests, dependencies and deployment environments.	Coordinate repository-scale changes across teams and services.	Integrate multimodal streams over extended workflows and environments.	Preserve project state across sessions, tools and collaborators.	Coordinate tools, APIs and environments across complex workflows.
III	Reason competently across a substantial professional domain.	Apply deep domain knowledge to professional tasks.	Build and debug substantial applications within one stack.	Own a feature across design, code, test and review.	Work fluently across several modalities in professional tasks.	Maintain context across long professional tasks and documents.	Use multiple tools reliably within a professional domain.
II	Complete several dependent reasoning steps in a defined task.	Use relevant knowledge within a defined workflow.	Implement a defined feature or function.	Complete a bounded ticket within an existing process.	Combine text with image, audio or structured data.	Retain relevant state across a bounded interaction.	Choose among tools in a bounded workflow.
I	Solve a bounded, immediate inference.	Retrieve narrow learned facts.	Generate or repair a bounded code fragment.	Assist a developer on a local task.	Interpret one additional modality in a bounded task.	Use immediate prompt context.	Call one predefined tool.

Periodic Table of AI Performance and Capability. Exhibit 18.2B

AI Level	Agentic Capability	Speed / Latency	Cost / Efficiency	Accuracy / Factuality	Instruction Following	Helpfulness / Task Effectiveness	Physical / Environmental Capability
VII	Sustain general autonomous action across global, physical & institutional environments.	Operate globally at machine speed across civilisation-scale decision & action loops.	Make abundant high-level cognitive and physical capability a near-ubiquitous utility.	Maintain globally robust truth-seeking under strategic manipulation & rapidly changing reality.	Act under legitimate civilisation-scale aims and constraints across jurisdictions and generations.	Increase physical and relational abundance at civilisation scale without displacing legitimate human aim.	Sustain general autonomous physical action across global environments and infrastructure.
VI	Coordinate multi-agent activity across organisations or national-scale systems.	Provide near-real-time analysis and action across large institutions and infrastructure.	Drive national-scale substitution, augmentation & optimisation.	Support high-stakes institutional decisions with evidence & systematic challenge.	Operate under institution-scale policy, law and public-interest constraints.	Support national-scale productivity, science, services & institutional problem solving.	Operate consequential physical systems across industries, logistics and critical infrastructure.
V	Pursue complex goals over long horizons, delegate to subagents and adapt plans.	Run many parallel high-value cognitive tasks faster than human teams.	Make autonomous or agentic cognitive work economical at very large scale.	Maintain calibrated truth-seeking across complex, adversarial and changing domains.	Balance multiple principals, changing constraints and delegated authority across agents.	Improve system-level outcomes over extended goals rather than merely local outputs.	Coordinate versatile robots, laboratories or infrastructure across complex real environments.
IV	Manage extended tasks, tools and recovery with limited supervision.	Coordinate complex multimodal and tool-using workflows with low delay.	Support large-scale deployment across organisations with manageable compute and supervision.	Cross-check multiple sources and recover from detected contradictions.	Maintain layered instructions, permissions & role constraints across long workflows.	Advance complex projects involving multiple constraints and stakeholders.	Operate across semi-structured workflows with perception, tools & recovery.
III	Pursue a defined professional goal with periodic human checkpoints.	Operate fast enough for professional interactive work at scale.	Deliver professional-grade work at much lower cost than human-only execution.	Produce reliable professional outputs with uncertainty signalling.	Interpret professional intent and resolve ordinary ambiguity.	Deliver professional-grade outcomes in a defined domain.	Perform repeatable robotic or laboratory tasks under close supervision.
II	Execute a short multi-step plan under close supervision.	Support real-time bounded workflows.	Make repeated bounded assistance economically practical.	Maintain acceptable accuracy in bounded workflows.	Follow multi-part instructions and local constraints.	Complete a bounded user task.	Execute bounded actions in structured environment
I	Take one bounded action after instruction.	Return a useful response within ordinary interactive delay.	Deliver useful output at high marginal cost.	Answer simple factual questions with basic checking.	Follow a direct bounded instruction.	Produce a locally useful answer.	Control one simple device or simulated actuator.

Where performance appears to sit now. The current reading is broad and provisional. It shows the general position of the US and Chinese frontier-AI ecosystems rather than ranking individual laboratories. The most important performance finding is convergence: both ecosystems sit in a similar frontier band across many cognitive capabilities.

Current Performance Reading: US and Chinese Frontier Ecosystems. Exhibit 18.3

Level	Reasoning	Knowledge	Coding	Software Engineering	Multimodal Capability	Context / Memory	Tool Use	Agentic Capability	Speed / Latency	Cost / Efficiency	Accuracy / Factuality	Instruction Following	Helpfulness / Task Effectiveness	Physical / Environmental Capability
VII	Sustain cross-domain, global and intergenerational reasoning about civilisation-shaping systems.	Integrate global knowledge continuously across disciplines, cultures and generations.	Design, build and adapt civilisation-scale digital infrastructure with sustained autonomy.	Continuously engineer and re-engineer global critical digital systems.	Maintain civilisation-scale situational understanding across digital and physical modalities.	Maintain long-duration, globally distributed memory while preserving provenance and correction.	Coordinate global digital and physical tool systems while remaining governable.	Sustain general autonomous action across global, physical & institutional environments.	Operate globally at machine speed across civilisation-scale decision & action loops.	Make abundant high-level cognitive and physical capability a near-ubiquitous utility.	Maintain globally robust truth-seeking under strategic manipulation & rapidly changing reality.	Act under legitimate civilisation-scale aims and constraints across jurisdictions and generations.	Increase physical and relational abundance at civilisation scale without displacing legitimate human aim.	Sustain general autonomous physical action across global environments and infrastructure.
VI	Reason across institutions, economies and strategic actors at national scale.	Support institution- or nation-scale knowledge systems and scientific discovery.	Operate across major software [illegible] physical systems.	Manage institution-scale software systems and interdependent infrastructure.	Fuse large-scale multimodal feeds across organisations and physical environments.	Maintain institution-scale operational memory with selective retrieval and revision.	Operate tool ecosystems spanning institutions, infrastructure and strategic domains.	Coordinate multi-agent activity across organisations or national-scale systems.	Provide near-real-time analysis and action across large institutions and infrastructure.	Drive national-scale [illegible]	Support high-stakes institutional decisions with evidence & systematic challenge.	Operate under institution-scale policy, law and public-interest constraints.	Support national-scale productivity, science, services & institutional problem solving.	Operate consequential physical systems across industries, logistics and critical infrastructure.
V	Model complex systems, agents [illegible]	Maintain broad cross-domain [illegible]	Plan and execute large [illegible] changes with tools and agents.	Run extended engineering [illegible] recovery.	Reason over persistent [illegible] tools as one system.	Use persistent memory across [illegible] multiple agents.	Discover, sequence and [illegible] tasks.	Pursue complex goals over long [illegible] adapt plans.	Run many parallel high-value cognitive [illegible] teams.	Make autonomous or agentic [illegible] very large scale.	Maintain calibrated truth-seeking [illegible] changing domains.	Balance multiple principals, [illegible] delegated authority across agents.	Improve system-level outcomes over [illegible] timely local outputs.	Coordinate versatile robots, laboratories or infrastructure across complex real environments.
IV	Integrate multiple domains, constraints and evidence sources.	Integrate knowledge across multiple domains and sources.	Work across repositories, tests, dependencies and deployment environments.	Coordinate repository-scale changes across teams and services	Integrate multimodal streams over extended workflows and environments.	Preserve project state across sessions, tools and collaborators.	Coordinate tools, APIs and environments across [illegible] workflows	Manage extended tasks, tools and recovery [illegible] limited supervision.	Coordinate complex multimodal and tool-using workflows with low delay.	Support large-scale deployment across organisations with manageable compute and supervision.	Cross-check multiple sources and [illegible]	Maintain layered instructions, permissions & role constraints across long workflows.	Advance complex projects involving multiple constraints and stakeholders.	Operate across semi-structured workflows with [illegible]
III	Reason competently across a substantial professional domain.	Apply deep domain knowledge to professional tasks.	Build and debug substantial applications within one stack.	Own a feature across design, code, test and review.	Work fluently across several modalities in professional tasks.	Maintain context across long professional tasks and documents.	Use multiple tools reliably within a professional domain.	Pursue a defined professional goal with periodic human checkpoints.	Operate fast enough for professional interactive work at scale.	Deliver professional-grade work at much lower cost than human-only execution.	Produce reliable professional outputs with uncertainty signalling.	Interpret professional intent and resolve ordinary ambiguity.	Deliver professional-grade outcomes in a defined domain.	Perform repeatable [illegible] supervision.

USA Emerging USA China Emerging China

Both ecosystems are concentrated around Level V, with selected Level VI characteristics emerging in reasoning, coding, efficiency and strategic domains. The picture is uneven. Cognitive and software capability has moved much further than general physical capability. That unevenness matters because the consequences change again when advanced reasoning is connected to robots, laboratories and infrastructure.

Reasoning, knowledge, multimodal capability and coding are already strong across both ecosystems. Agentic capability, long-duration memory, reliable factuality and general physical action remain less settled. Performance is therefore high, but not uniform across every noun.

4. Safety: The Nouns

Safety should now be examined separately. The safety nouns describe the controls needed around increasingly capable systems. They range from refusal and robustness through to monitoring, human agency, access to power and the ability to learn from failure.

Safety Nouns. Exhibit 18.4

Noun	Meaning
Refusal / Constraint	The ability to decline, limit or safely redirect actions that fall outside an authorised or acceptable operating boundary.
Bias / Fairness	The ability to recognise and reduce unjustified differential treatment while making relevant trade-offs explicit.
Cyber Safety	The controls that prevent powerful cyber capability from being used to discover, exploit, automate or scale harmful intrusion.
Biological / Chemical Safety	The controls that limit AI assistance that could materially enable hazardous biological or chemical design, development or deployment.
Jailbreak Resistance	The robustness of safeguards against deliberate attempts to circumvent system rules through adversarial prompting or multi-turn attack.
Agentic Safety	The ability to keep increasingly autonomous systems inside intended aims, boundaries, permissions and escalation rules.
Monitoring / Controllability	The ability for humans or supervising systems to observe, interrupt, constrain, redirect and stop AI behaviour.
Deception / Manipulation Control	The ability to detect and limit strategically misleading, covertly persuasive or manipulative behaviour by AI systems.
Robustness / Reliability	The ability to remain within acceptable behaviour under unfamiliar, adversarial, degraded or rapidly changing conditions.
Access / Power Control	The mechanisms that determine who can obtain consequential capabilities, tools, permissions, compute and downstream reach.
Human Agency	The extent to which humans retain meaningful decision, challenge, override and responsibility rather than being displaced by automation.
Learning / Correction	The ability of the safety system to update thresholds, safeguards, permissions and governance after incidents, evaluations or new capabilities reveal weaknesses.

These nouns are not claims that safety is finished or complete. They are a practical set of recurring control functions that can be observed as the boundary of consequence grows.

5. Safety: The Periodic Table and Current Position

The Safety Periodic Table follows a different path from performance. Higher levels are not simply “safer”. The boundary of control becomes wider. Levels I to IV sit mainly inside products, workflows and laboratories. Level V requires national public authority. Level VI requires international coordination. Level VII requires global or civilisational control. The test is whether the control exists at the level of the consequence being created.

Safety Periodic Table. Exhibit 18.5A

Level	Refusal / Constraint	Bias / Fairness	Cyber Safety	Biological / Chemical Safety	Jailbreak Resistance	Agentic Safety
VII	Maintain globally legitimate constraints on civilisation-threatening capabilities and uses.	Continually renegotiate global fairness across cultures, generations and unequal power.	Coordinate globally against AI-enabled systemic cyber conflict and infrastructure collapse.	Maintain global and intergenerational control over AI-enabled catastrophic biological or chemical risk.	Maintain adaptive global defences against rapidly evolving, machine-generated attacks on safeguards.	Preserve human civilisational control over highly autonomous, self-extending or strategically capable systems.
VI	Coordinate compatible constraints, thresholds and enforcement across major countries and frontier ecosystems.	Coordinate rights and fairness standards across jurisdictions where AI systems and effects cross borders.	Coordinate cross-border cyber controls, threat intelligence, incident response and minimum safeguards.	Coordinate international biosecurity thresholds, laboratory controls, detection and response.	Share attack intelligence, evaluation methods and minimum safeguards across major AI jurisdictions.	Coordinate international rules, testing and stop conditions for agents able to act across borders.
V	Apply enforceable national constraints to high-consequence capabilities across firms and infrastructure.	Embed fairness and rights protections in legitimate national institutions, standards and remedies.	Place high-consequence cyber-AI capability under binding national security and public-safety control.	Apply binding national biosecurity standards, access controls and incident response.	Require independent national evaluation and minimum robustness standards for high-consequence systems.	Require binding public controls for autonomous systems capable of national-scale consequence.
IV	Tie frontier-model permissions and deployment gates to organisational risk thresholds.	Govern fairness trade-offs through documented organisational testing and challenge.	Evaluate frontier cyber capability, restrict access and gate deployment against thresholds.	Evaluate frontier bio/chemical capability, restrict access and trigger mitigations at thresholds.	Use systematic red-teaming, adversarial evaluation and mitigation before frontier release.	Apply frontier-agent evaluations, sandboxes, escalation rules and stop conditions inside the lab.
III	Apply consistent product-level policy across recurring users and contexts.	Measure and mitigate recurring disparities across a deployed product.	Monitor and limit cyber misuse across a deployed model or service.	Monitor and filter hazardous capability across products and user populations.	Maintain safeguards across diverse users, languages and attack patterns.	Bound product agents with tool permissions, logs, checkpoints and recovery.
II	Constrain a bounded workflow and escalate exceptions.	Apply basic fairness rules within a workflow.	Constrain tool-assisted cyber work in supervised workflows.	Constrain sensitive assistance in supervised workflows.	Resist common multi-turn attempts within a bounded workflow.	Supervise a short plan and require approval for sensitive steps.
I	Block an explicitly prohibited output.	Avoid obvious discriminatory output in a bounded case.	Refuse plainly malicious cyber instructions.	Refuse directly harmful bio/chemical requests.	Resist simple adversarial prompting.	Keep a single action inside explicit permissions.

Safety Periodic Table. Exhibit 18.5B

Level	Monitoring / Controllability	Deception / Manipulation Control	Robustness / Reliability	Access / Power Control	Human Agency	Learning / Correction
VII	Maintain global visibility and effective intervention against civilisation-scale loss of control.	Maintain global institutions able to detect and respond to strategically deceptive superhuman systems.	Preserve civilisational resilience despite correlated failures, strategic attack or transformative system change.	Prevent irreversible concentration or uncontrolled diffusion of civilisation-shaping AI power globally.	Ensure humanity retains ultimate authority over civilisation-scale aims, boundaries and settlements.	Continuously recalibrate global safety architecture while preserving memory of previous failures and settlements.
VI	Build cross-border visibility, incident sharing and coordinated intervention for internationally consequential systems.	Coordinate international detection, disclosure and response to strategic deception and manipulation.	Coordinate international resilience standards for interconnected AI, infrastructure and supply chains.	Coordinate international controls over frontier compute, model access, supply chains and dangerous diffusion.	Preserve meaningful human authority across international institutions and cross-border systems shaped by AI.	Learn across countries and update international standards, agreements and joint response mechanisms.
V	Give national public authorities inspection, testing, intervention and override powers over high-consequence AI.	Require independent national testing, incident disclosure and enforceable controls for deceptive capability.	Set and enforce national reliability standards for AI embedded in essential services and infrastructure.	Use legitimate national authority to govern high-consequence compute, models, infrastructure and strategic access.	Protect national democratic, legal or equivalent public authority against de facto AI substitution.	Update national law, standards and institutional capacity as capability and risk evolve.
IV	Maintain standing sensing, thresholds, decision rights and stop mechanisms for frontier systems.	Evaluate frontier systems for scheming, concealment, reward hacking and manipulation under pressure.	Stress-test frontier systems under adversarial, degraded and novel conditions before release.	Govern frontier weights, compute, privileged tools, insiders and deployment authority inside the lab.	Design organisational controls so AI cannot silently displace accountable human decision rights.	Revise frontier evaluations, thresholds, governance and mitigations as evidence changes.
III	Log, investigate and correct behaviour across a deployed product.	Test and mitigate recurring persuasion, impersonation and strategic deception in products.	Maintain acceptable behaviour across diverse users and deployment conditions.	Control product tiers, identity, tools and sensitive capabilities across user populations.	Preserve user choice, appeal and override across deployed products.	Use incidents and user evidence to recalibrate product safeguards.
II	Monitor a workflow and hand control back to a human.	Detect misleading behaviour within bounded interactions.	Recover safely from routine workflow errors and edge cases.	Apply role-based permissions in a workflow.	Require meaningful approval at important workflow decisions.	Update workflow rules after recurring errors.
I	Observe and interrupt one output or action.	Block obvious deceptive or manipulative output.	Remain stable under ordinary prompt variation.	Restrict one sensitive feature or tool.	Keep a human able to accept or reject an immediate output.	Patch an immediate failure.

Where safety appears to sit now - Safety is harder to read than performance because public evidence is less complete and the US and Chinese systems organise control differently. The United States relying on developer frameworks, external evaluation and developing public rules. China preferring administrative and platform controls.

Current Safety Reading: US and Chinese Frontier Ecosystems. Exhibit 18.6

	Refusal / Constraint	Bias / Fairness	Cyber Safety	Biological / Chemical Safety	Jailbreak Resistance	Agentic Safety	Monitoring / Controllability	Deception / Manipulation Control	Robustness / Reliability	Access / Power Control	Human Agency	Learning / Correction
VII	Maintain globally legitimate constraints on civilisation-threatening capabilities and uses.	Continually renegotiate global fairness across cultures, generations and unequal power.	Coordinate globally against AI-enabled systemic cyber conflict and infrastructure collapse.	Maintain global and intergenerational control over AI-enabled catastrophic biological or chemical risk.	Maintain adaptive global defences against rapidly evolving, machine-generated attacks on safeguards.	Preserve human civilisational control over highly autonomous, self-extending or strategically capable systems.	Maintain global visibility and effective intervention against civilisation-scale loss of control.	Maintain global institutions able to detect and respond to strategically deceptive superhuman systems.	Preserve civilisational resilience despite correlated failures, strategic attack or transformative system change.	Prevent irreversible concentration or uncontrolled diffusion of civilisation-shaping AI power globally.	Ensure humanity retains ultimate authority over civilisation-scale aims, boundaries and settlements.	Continuously recalibrate global safety architecture while preserving memory of previous failures and settlements.
VI	Coordinate compatible constraints, thresholds and enforcement across major countries and frontier ecosystems.	Coordinate rights and fairness standards across jurisdictions where AI systems and effects cross borders.	Coordinate cross-border cyber controls, threat intelligence, incident response and minimum safeguards.	Coordinate international biosecurity thresholds, laboratory controls, detection and response.	Share attack intelligence, evaluation methods and minimum safeguards across major AI jurisdictions.	Coordinate international rules, testing and stop conditions for agents able to act across borders.	Build cross-border visibility, incident sharing and coordinated intervention for internationally consequential systems.	Coordinate international detection, disclosure and response to strategic deception and manipulation.	Coordinate international resilience standards for interconnected AI, infrastructure and supply chains.	Coordinate international controls over frontier compute, model access, supply chains and dangerous diffusion.	Preserve meaningful human authority across international institutions and cross-border systems shaped by AI.	Learn across countries and update international standards, agreements and joint response mechanisms.
V	Apply enforceable national constraints to high-consequence [illegible]	Embed fairness and rights protections in legitimate national institutions, standards and remedies.	Place high-consequence cyber-AI capability under national [illegible] and public-sector control.	Apply binding national biosecurity standards, access [illegible] and incident	Require independent national evaluation and minimum robustness standards for high-consequence systems.	Require binding public controls for autonomous systems capable of national-scale consequence.	Give national public authorities inspection, testing, [illegible] powers over high-consequence AI.	Require independent national testing, incident disclosure and enforceable controls for deceptive capability.	Set and enforce national reliability standards for AI embedded in essential services and infrastructure.	Use legitimate national authority to [illegible] high-[illegible] infrastructure and strategic access.	Protect national democratic, legal or equivalent public authority against de facto AI substitution.	Update national law, standards and [illegible] capability
IV	Tie frontier-model permissions and [illegible]	Govern fairness trade-offs through [illegible]	Evaluate frontier cyber capability, [illegible]	Evaluate frontier bio/chemical [illegible] thresholds.	Use systematic red-teaming, adversarial [illegible]	Apply frontier-agent evaluations, [illegible] inside the lab.	Maintain standing sensing, thresholds, [illegible]	Evaluate frontier systems for [illegible] pressure.	Stress-test frontier systems under [illegible] release.	Govern frontier weights, compute, [illegible] authority inside the lab.	Design organisational controls so AI cannot [illegible]	Revise frontier evaluations, [illegible] evidence changes.
III	Apply consistent product-level policy across recurring users and contexts.	Measure and mitigate recurring disparities across a deployed product.	Monitor and limit cyber misuse across a deployed model or service.	Monitor and filter hazardous [illegible] across [illegible] user populations.	Maintain safeguards [illegible]	Bound product [illegible] recovery.	Log, investigate and correct behaviour across a deployed product.	Test and mitigate [illegible] in products.	Maintain acceptable behaviour across diverse users and deployment conditions.	Control product tiers, identity, tools and sensitive capabilities across user populations.	Preserve user choice, appeal and override across deployed products.	Use incidents and user evidence to recalibrate product safeguards.
II	Constrain a bounded workflow and escalate exceptions.	Apply basic fairness rules within a workflow.	Constrain tool-assisted cyber work in supervised workflows.	Constrain sensitive assistance in supervised workflows.	Resist common multi-turn attempts within a bounded workflow.	Supervise a short plan and require approval for sensitive steps.	Monitor a workflow and hand control back to a human.	Detect misleading behaviour within bounded interactions.	Recover safely from routine workflow errors and edge cases.	Apply role-based permissions in a workflow.	Require meaningful approval at important workflow decisions.	Update workflow rules after recurring errors.
I	Block an explicitly prohibited output.	Avoid obvious discriminatory output in a bounded case.	Refuse plainly malicious cyber instructions.	Refuse directly harmful bio/chemical requests.	Resist simple adversarial prompting.	Keep a single action inside explicit permissions.	Observe and interrupt one output or action.	Block obvious deceptive or manipulative output.	Remain stable under ordinary prompt variation.	Restrict one sensitive feature or tool.	Keep a human able to accept or reject an immediate output.	Patch an immediate failure.

USA Emerging USA China Emerging China

The broad safety picture sits lower than the performance picture. Many controls remain around Levels III to V. Product constraints and organisational safeguards are real, but several important areas - agentic safety, jailbreak resistance, deception control and monitoring - are still developing while capability continues to move.

The more important point is the level boundary. A Level V safety arrangement is a national arrangement. That may be necessary, but it is not enough for risks that cross borders or could affect civilisation more broadly. **Those safety requirements are Level VI international coordination and, at the highest consequence, Level VII global control**.

6. The Gap Between Performance and Safety

Only now should performance and safety be put beside each other. The current picture suggests that frontier performance is concentrated around Level V, with some movement toward Level VI, while several critical safety functions remain around Levels III to V.

Three gaps stand out. Agentic capability is moving faster than agentic safety and monitoring. Cyber and biological capability can create consequences beyond the laboratory or even the country while much practical control remains organisational or national. Physical capability is still below cognitive capability, but the risk can change quickly when advanced reasoning is connected to machines, laboratories and infrastructure.

This does not prove that frontier AI is unsafe. It shows where attention is required. The key question is whether capability is moving into a level for which the matching control does not yet exist. If it is, the gap is an early warning.

The US-China comparison from Chapter 18 makes the point sharper. Both countries can improve national safety. But national safety alone cannot settle risks that are international or global. Being satisfied with Level V control is therefore not an adequate end point when the consequence has moved to Levels VI or VII.

7. What This Means for People Working on AI

People working on AI need to understand more than AI. They need to understand their own brains and the institutions around them. The humans building, releasing and governing AI are themselves part of the safety system.

The same brain that can build extraordinary technology is also shaped by threat, status, competition, identity, felt truths and bias. Under pressure, people can narrow their thinking, defend existing beliefs, discount longer-term consequences and treat disagreement as threat. Technical ability does not remove those tendencies. In some settings it can simply make people better at defending them.

Self-observation matters. So do restraint, truth-seeking and the ability to hold competing goods rather than collapsing them into one objective. A technically brilliant person or team that does not understand its own human behaviour can still make poor decisions about capability, release, access and safety.

The second requirement is to understand institutional architecture. Decisions are shaped by incentives, authority, information flows, competition, challenge, accountability and the level at which responsibility sits. Strong people can still produce poor outcomes if the institution pushes capability, speed or commercial advantage harder than challenge, correction and wider responsibility.

This is where the pilots should start to be used inside AI laboratories. Primitives literacy can give teams a common language. Personal mirrors can help people see their own predispositions and reactions. Group mirrors can make team behaviour and blind spots visible. Periodic Tables can show the level at which capability and control are operating. The Abacus can expose competing goods. Control-and-correction loops can strengthen sensing, challenge and learning. Truth-verification processes can improve how evidence is tested before consequential decisions are made.

These tools should not sit outside the laboratories as ideas for later. They can begin to be tested now. The purpose is not to slow technical work for its own sake. It is to improve the human and institutional systems making decisions about increasingly powerful technology.

8. What This Means for AI

Increasing intelligence should not be treated as one simple upward curve. Capability is made up of different things, and the consequence of capability depends on what the system can reach, how long it can act, which tools it can use and whether its actions can be observed and corrected.

As reasoning, agency, tool use and physical action move upward, control has to move with them and sometimes above them. An AI that acts across an ecosystem cannot be controlled only by output filtering. An AI that can create national-scale consequences cannot be governed only by a private laboratory. An AI capable of international or civilisation-scale consequence needs control at those higher levels.

The Periodic Tables make that movement visible. They do not solve the safety problem. They show where performance is moving, where control is operating and where the two are separating. That visibility is the starting point for correction.