Chapter 5. Cybernetic control systems for machines can be applied to humans

Introduction

I have spent much of my working life inside very large physical systems: mines, smelters, manufacturing plants and gas fields. None of them can operate safely or productively by relying on goodwill, intelligence or human supervision alone. They depend on interlinked systems that continuously sense what is happening, compare it with what should be happening, correct deviation and learn from repeated performance.

That is why this chapter starts with machines. Mechanical and industrial systems make cybernetics visible. A conveyor either starts or does not. A valve opens or closes. Pressure, temperature, flow and speed sit inside a permitted range or move outside it. Sensors detect the actual state. Controllers compare it with the intended state. Actuators respond. Higher-order systems coordinate multiple local loops and bring human operators into the picture when the disturbance exceeds local capability.

Once that architecture is clear, we can make one controlled move at a time. First, we establish the principles in physical and mechanical systems. Second, we apply the same grammar to human systems, where aims, evidence, authority and correction become more contested. Third, we introduce AI as a new capability inside human control and correction systems. Fourth, we turn the same principles back onto AI itself and ask how an increasingly capable AI system must be bounded, observed and corrected.

The central proposition is simple: the cybernetic grammar remains remarkably stable, but the difficulty of applying it rises sharply as we move from machines, to humans, to AI-enabled human systems, and finally to AI itself.

The nine recurring components are Aim/Purpose, Boundary, Range, Sensing, Comparison, Interpretation, Decision Rule, Correction/Action, and Learning/Recalibration. The nouns remain substantially the same. What changes are the verbs: what must be sensed, how evidence is interpreted, who has authority to act, how quickly correction occurs, and who corrects the controller.

The chapter therefore builds the argument in four separate stages. That separation matters. If we mix machines, people and AI too early, the common architecture becomes harder to see and the important differences disappear. We will do the opposite: make the mechanical principles explicit first, then add human complexity,

then add AI as a capability, and only then address AI as an object of control and correction.

How this chapter unfolds

1. Mechanical systems make cybernetics visible — Mechanical systems make cybernetics visible because their aims, boundaries, sensing, comparison and correction can be seen directly.

2. Human systems use the same cybernetic grammar — Human systems use the same cybernetic grammar, but human judgement, competing aims, power and partial rationality make correction harder.

3. AI can strengthen human control and correction systems — AI can strengthen human control and correction systems by making them more observable, faster and more capable of learning.

4. The same cybernetic principles must control AI itself — The same cybernetic principles must also be applied to AI itself, so that AI remains inside human-set aims, boundaries and legitimate correction.

5. What this means for humans — Human systems need clear aims, boundaries, accountability and higher-level correction, while well-designed rules can create more freedom by making safe local action possible.

6. What this means for AI — AI can strengthen human correction systems and operate with substantial freedom inside clear limits, but it must remain visible, bounded, independently challengeable and correctable.

1. Mechanical systems make cybernetics visible

Mechanical systems make cybernetics visible because their aims, boundaries, sensing, comparison and correction can be seen directly.

The advantage of beginning with mechanical systems is precision. We can see the variable being controlled, the permitted range, the sensor, the comparison, the decision rule and the physical response. The system can still be complex, but its control logic is comparatively explicit. That gives us a clean base before we introduce human judgement, legitimacy or AI.

Every control and correction system has nine recurring components

At its most compressed, a control and correction system can be described through nine stable nouns. These are not nine sequential boxes that must always occur in a

neat line. They are the recurring functions that have to exist somewhere in a viable correction system.

Nine standard nouns of control and correction systems. Exhibit 5.1

Step	Standard noun	Meaning
1	Aim / Purpose	Define what the system is trying to maintain, improve, or become.
2	Boundary	Define what is inside the system, what is outside it, and what must be regulated.
3	Range	Specify the target condition, acceptable range, or performance standard.
4	Sensing	Observe what is actually happening in the system and its environment.
5	Comparison	Compare actual state with desired state and identify variance.
6	Interpretation	Diagnose what the variance means and why it is occurring.
7	Decision Rule	Decide what kind of response is required, by whom, and at what level.
8	Correction / Action	Intervene to reduce harmful variance or amplify useful variation.
9	Learning / Recalibration	Update the aim, standard, sensing, rules, or structure if the system itself has changed, using the idea of levels seen in chapter 3.

A thermostat is the simplest illustration. The aim is comfort. The boundary is the room. The range is the acceptable temperature band. A sensor observes the present temperature. Comparison identifies the difference between actual and desired condition. Interpretation determines what that variance means. A decision rule determines whether action is required. The heater acts. The resulting temperature becomes new evidence, and repeated failure forces recalibration of the settings, sensor, heater or even the desired range.

The important point is not the thermostat. It is the architecture. Once the nine components are visible in a simple loop, we can watch the same grammar reappear in more complicated systems.

The nine components remain, while their verbs become more complex

As physical systems increase in scale, the nine nouns do not disappear. A local controller may regulate one variable. A machine controller coordinates several. A plant-level system integrates many machines. A network-level system has to anticipate interactions across multiple plants or physical networks. The nouns remain; the work inside them becomes more demanding.

At higher mechanical levels, sensing becomes aggregation of distributed telemetry. Comparison becomes reconciliation against multiple constraints. Interpretation

becomes diagnosis of interactions and cascading effects. Decision rules become contingent and system-wide. Correction becomes reconfiguration rather than a simple on/off response. Learning becomes model and architecture recalibration.

Exhibit 5.2 makes one principle visible: increasing level does not require a new theory of cybernetics. It requires more capable verbs inside the same control and correction grammar.

The nine nouns of control and correction systems and their levelled verbs applied to mechanical systems. Exhibit 5.2

Level	Aim / Purpose	Boundary	Range	Sensing	Comparison	Interpretation	Decision Rule	Correction / Action	Learning / Recalibration
VII	Orchestrate system-of-systems performance towards an externally defined purpose	Couple interacting physical networks while preserving their operating limits	Balance resilience envelopes across multiple connected systems	Synthesise system-wide telemetry, forecasts and external conditions	Simulate actual and projected states against multi-system constraints	Anticipate cascading, emergent and cross-system effects	Coordinate contingent responses across interacting control domains	Reconfigure flows, capacities and operating modes across systems	Recalibrate models, architecture and control parameters under human authority
VI	Integrate multiple major systems towards common operating outcomes	Federate subsystem boundaries, interfaces and dependencies	Differentiate operating envelopes across assets, regions and conditions	Fuse distributed signals from multiple control systems	Reconcile network performance against capacity, reliability and constraint requirements	Diagnose interactions and propagation across interconnected systems	Dispatch responses according to system-wide priorities and contingencies	Reroute loads, flows or capacity to preserve overall performance	Adapt supervisory models, rules and parameters from accumulated system performance
V	Optimise whole-plant or whole-system performance	Define functional interfaces among major subsystems	Allocate operating envelopes and tolerances across interconnected units	Supervise distributed sensing across the whole system	Benchmark subsystem and total-system performance against plant requirements	Model cross-functional causes, constraints and interactions	Arbitrate competing control demands according to plant priorities	Redistribute loads, resources and operating conditions across subsystems	Reconfigure control strategies and architecture after sustained or systemic variance
IV	Coordinate an integrated process, line or operating unit	Partition interacting equipment and process zones	Balance multiple operating conditions within a shared process envelope	Aggregate signals from interconnected machines and process stages	Reconcile interacting variances across the process	Trace variance through causal relationships among components	Prioritise responses where several control demands interact	Coordinate actuators and process changes across the operating unit	Retune process logic, sequences and control relationships from performance data
III	Maintain the performance of a machine or equipment assembly	Delimit equipment, interfaces and controlled variables	Regulate several related variables within specified tolerances	Monitor multiple operating signals and equipment states	Correlate measured conditions with specified operating states	Diagnose identifiable equipment or process faults	Sequence programmed responses according to defined logic	Compensate for disturbance through coordinated actuator response	Tune settings, thresholds or control parameters from operating history
II	Hold a simple physical condition or output	Contain the controlled device, variable or immediate space	Set a fixed target, limit or acceptable tolerance	Detect the immediate state of the controlled variable	Compare the measured state with the preset condition	Classify the variance according to predetermined logic	Trigger a predefined response when a threshold is crossed	Adjust the controlled variable directly	Recalibrate the setting or sensor through preset or external intervention

Design constraints and operational constraints solve different problems

Mechanical systems also reveal a distinction that becomes important later. Design constraints define the operating envelope: what the system is built to do, what it must not do, and the physical conditions it must never exceed. Operational constraints govern behaviour inside that envelope: setpoints, sequences, alarms, interlocks, permissions, overrides and shutdowns.

Both are necessary. Design constraints without operational translation remain statements of intent. Operational constraints without sound design constraints can make the wrong objective highly efficient. A tightly controlled process can still damage the larger plant if its purpose, boundary or operating envelope is poorly set.

Constraint for mechanical control and correction systems. Exhibit 5.3

Constraint type	Function	Source	Typical mechanisms
Design constraints - Engineering constraints that define what the mechanical system is designed and permitted to do. They establish its physical configuration, capacities, operating envelope, interfaces, safe limits and no-go conditions within which operation must occur.	Define the control envelope within which the mechanical system is designed to operate. Establish its purpose, physical boundaries, capacities, safe limits, allowable ranges and conditions that must not be exceeded.	System and equipment design; engineering specifications; equipment ratings; process requirements; safety and environmental design limits; interface requirements with connected systems.	Design capacity, pressure and temperature limits, equipment ratings, physical guards, relief systems, engineered safety margins, hard interlocks, trip limits, containment, fail-safe design and shutdown architecture.
Operational constraints - Control constraints that govern how the mechanical system operates within its design envelope. They regulate sequence, setpoints, timing, variation and corrective response so that the system remains within its designed limits and required performance range.	Control how the system behaves within the design envelope. Translate design limits into real-time operating rules, sequences, setpoints and corrective responses that keep performance within the permitted range.	Control logic; operating setpoints; process conditions; sensor feedback; equipment status; programmed sequences; upstream and downstream operating requirements.	PLC logic, control algorithms, setpoints, alarms, permissions, interlocks, sequencing, actuator limits, feedback loops, automatic trips, overrides and controlled shutdowns.

Large physical systems need both local and higher-order correction

A mine, refinery or factory is never one loop. It is a network of local loops connected horizontally along the flow of work and vertically into higher levels of supervision. The output of one system becomes the input, constraint or disturbance for another. Local optimisation is therefore not enough.

Horizontal correction is concerned with interfaces. A machine must not only meet its own target; its output must be usable by the next stage. Quality, quantity, timing, cost, information integrity and reliability become shared variables. A machine that maximises throughput while producing material outside downstream limits is not helping the plant.

Vertical correction deals with a different problem. Supervisory control and data acquisition systems (SCADA) gather information from multiple local controllers, make the wider process visible, manage alarms and support intervention when local loops interact, conflict or fail. The higher level does not need to take over every lower-level task. Its value comes from seeing what the local system cannot see and correcting what the local system cannot correct.

This creates the first form of L+1 logic. A local system should correct what lies within its competence and authority. A higher-level system sets broader constraints, monitors interactions and intervenes when the disturbance exceeds local range, crosses boundaries or exposes a flaw in the governing design. Downward movement carries purpose, limits and authority. Upward movement carries evidence, variance, exception and learning.

Vertical and horizontal correction in an industrial system. Exhibit 5.4

Layer	Primary role	What it sees	When it intervenes
Local controller / PLC	Regulates a bounded machine or process	Immediate variables, equipment state, local alarms	When the local variable or sequence leaves its permitted range
Horizontal interface	Protects hand-offs between connected systems	Quality, quantity, timing, reliability and compatibility of outputs/inputs	When one subsystem imposes a disturbance on another
Supervisory / SCADA	Coordinates multiple local loops	Patterns, interactions, common constraints and system-level alarms	When local loops conflict, interact or cannot correct the disturbance
Human apex	Sets purpose, design envelope and non-negotiable limits	The wider plant, safety, environment and external consequences	When the control architecture itself requires redesign or shutdown

Mechanical systems therefore give us a complete starting grammar: clear aims, explicit boundaries, defined ranges, sensing, comparison, interpretation, decision rules, action, learning, local correction, higher-order correction and interface management. Only now is it useful to ask what happens when the system being corrected contains human beings.

2. Human systems use the same cybernetic grammar

Human systems use the same cybernetic grammar, but human judgement, competing aims, power and partial rationality make correction harder.

Human systems are not machines. People interpret rules, defend identities, pursue status, cooperate, compete, conceal, reciprocate, retaliate and disagree about the aim itself. Yet human systems still have to solve the same basic control problem: define what they are trying to achieve, establish boundaries, sense reality, compare reality with expectation, interpret variance, decide, act and learn.

The shift from machines to humans therefore does not replace the mechanical grammar. It adds a much harder operating environment around it.

The same nine nouns now require human judgement and legitimacy

At lower human levels, aims, boundaries, expectations, sensing and correction can be immediate and personal. At higher levels, they operate across larger groups, institutions and societies, longer time horizons and more complex networks of interdependence. Aims become broader and more contested. Boundaries become legal and moral as well as physical. Sensing depends on distributed and independent evidence. Interpretation requires judgement. Decision rules require legitimate authority. Correction can mean changing behaviour, resources, institutions or the rules themselves.

The nine nouns of control and correction systems and their levelled verbs applied to human systems. Exhibit 5.5

Level	Aim / Purpose	Boundary	Range	Sensing	Comparison	Interpretation	Decision Rule	Correction / Action	Learning / Recalibration
VII	Steward human and planetary flourishing	Harmonise sovereignty with planetary interdependence	Universalise minimum standards and tolerances	Monitor planetary conditions and systemic risks	Reconcile outcomes with shared human aims	Deliberate across cultures, evidence and consequences	Constitutionalise legitimate rules for intervention	Rebalance systems before cascading failure spreads	Re-architect institutions through civilisational learning
VI	Integrate plural aims into common direction	Balance rights, jurisdictions and mutual obligations	Differentiate standards across legitimate contexts	Diagnose patterns from multiple evidence sources	Weigh competing outcomes, rights and risks	Contextualise variance across systems and groups	Adjudicate through legitimate plural deliberation	Remedy harms while preserving rights and legitimacy	Redesign arrangements through reflective review
V	Govern towards lawful national purposes	Legalise jurisdiction, rights and responsibilities	Codify standards, thresholds and tolerances	Audit performance, compliance and emerging risks	Evaluate outcomes against statutory standards	Reason from evidence, law and precedent	Legislate proportionate and lawful responses	Enforce compliance and remedy breaches	Reform institutions after systemic failure
IV	Command effort towards declared purpose	Decree territory, rank and permitted reach	Impose required conditions and limits	Inspect output, obedience and visible disorder	Judge performance against decreed requirements	Proclaim causes through authorised doctrine	Authorise through concentrated ruling power	Punish deviation and restore order	Reset commands after governing failure
III	Preserve group identity, continuity and survival	Mark membership, territory and reciprocal obligations	Customise norms to group conditions	Watch conduct, loyalty and changing conditions	Match conduct against shared custom	Explain variance through group narratives	Decide through recognised group authority	Restore conformity, trust and group cohesion	Realign custom through shared experience
II	Protect immediate family wellbeing and continuity	Enclose household, roles and responsibilities	Instruct simple limits for safe conduct	Notice immediate needs, risks and changes	Check conduct against known expectations	Sense causes through lived familiarity	Permit or prohibit immediate responses	Correct behaviour directly and promptly	Adjust routines through repeated experience

The mechanical and human exhibits therefore show both similarity and difference. Vertically, the nine nouns remain stable while their verbs become more complex. Horizontally, the verbs differ because the object of correction differs. Physical systems confront variation, coupling and technical complexity. Human systems confront all of that plus autonomy, conflicting interests, legitimacy, power and contested truth.

Human systems need both design constraints and operational constraints

The distinction between design and operational constraints carries across directly. In human systems, design constraints define what the system may be, do or pursue. Constitutions, laws, board mandates, regulatory limits, rights, prohibitions and social licence define the envelope. Operational constraints translate that envelope into roles, routines, measures, thresholds, approvals, audits, escalation and stop-work authority.

The difference is that human design constraints are not merely engineering choices. They arise from truth, felt truth, law, governance, deliberation and legitimate authority. That makes the design layer itself contestable and therefore in need of correction.

Two types of constraint needed for human adaptive systems. Exhibit 5.6

Constraint type	Function	Source	Typical mechanisms
Design constraints — Higher-level rules defining what the system may be, do or pursue. They establish the operating envelope, including how enduring paradoxes are bounded and legitimately balanced. They originate principally at Level L+1.	Convert truth, felt truth and paradox settlement into binding architecture. Establish the legitimate balance between competing goods, define non-negotiable limits and determine which trade-offs may be made at lower levels.	Higher-level truth, felt truth, deliberation, paradox settlement, law, governance and legitimate institutional authority.	Constitutions, laws, board mandates, regulatory standards, design rules, ethical limits, rights, prohibitions, licensing, liability and social licence.
Operational constraints — Rules governing how the system performs work, manages variation and corrects deviation within its design envelope. In human adaptive systems, they also guide the practical management of competing aims and paradoxical tensions.	Translate design constraints into execution, monitoring, correction and escalation. Enable bounded trade-offs, reveal when one side of a paradox dominates, restore balance and escalate when the underlying settlement requires reconsideration.	Management systems, engineering and process control and correction, operating judgement, local evidence, cybernetic feedback, delegated authority and the higher-level paradox settlement established at Level L+1.	SOPs, checklists, sensors, dashboards, decision rights, thresholds, approvals, quality assurance, audits, reviews, incident reports, escalation rules, stop-work authority and shutdown triggers.

Humans create an additional correction burden

Physical systems need control and correction because they are exposed to disturbance. Human systems need all of that and more. Human beings are powerful, adaptive, social, status-sensitive, emotionally animated and only partly rational. We can cooperate at extraordinary scale, but we can also defect, dominate, rationalise, panic, free-ride, protect our group and pursue short-term advantage at the expense of the whole.

The problem is not that people are bad. It is that human capability outruns reliable private self-restraint. The same fear that protects survival can become panic and scapegoating. The same ambition that produces invention can become predation. The same group loyalty that creates solidarity can become exclusion. The same intelligence that solves problems can also justify self-interest and hide failure.

Human characteristics and their control and correction requirements. Exhibit 5.7

Human condition	What it creates	Control system required	Risk to society if unrestrained
Survival pressure	Fear, urgency, defensive action	Safety rules, authority, escalation	Panic, violence, scapegoating and retreat into primitive protection behaviour
Desire and ambition	Energy, innovation, overreach	Boundaries, incentives, accountability	Exploitation, corruption, status competition and predatory accumulation
Status-seeking	Hierarchy, rivalry, domination	Role clarity, law, governance	Power becomes personalised; hierarchy becomes domination rather than ordered responsibility
Group loyalty / belonging attachment	Solidarity, obligation, exclusion, favouritism	General rules, rights, due process, impartial review	Society fragments into hostile belonging groups; truth becomes group-defensive and cooperation collapses across boundaries
Local optimisation	Functional success, system damage	Higher-level coordination and constraint	Each part maximises itself against the whole; the system becomes incoherent and self-damaging
Limited horizon	Short-term decisions, delayed harm	Calendar, planning, review, stewardship	Long-term costs accumulate invisibly until they appear as crisis, distrust or collapse
Rationalisation	Self-justifying error	Mirrors, audits, truth-seeking	False stories protect bad behaviour; reality is excluded until correction becomes violent or chaotic
Controller corruption	Capture, abuse, illegitimacy	L+1 control: guards above guards	Those with control use control for themselves; legitimacy fails and revolt pressure accumulates
Competing goods	Paradox: freedom/order, merit/equality, innovation/safety	Governance, deliberation, institutional design	One pole dominates; the suppressed pole returns as backlash, instability or systemic correction

Responsibility is not enough; human systems need accountability

This is where a specifically human distinction becomes important. Responsibility identifies who is expected to perform the work. Accountability connects that responsibility to observability, comparison, consequence and correction. A person can be responsible on paper while the system remains incapable of seeing whether the work was done, whether the result was acceptable or what happens when it was not.

A human control and correction system therefore needs teeth. That does not mean punishment as the default. It means that boundaries, authority, escalation and consequences must be real enough to change behaviour when persuasion, goodwill or local correction fail.

Human correction systems are recursive and every controller must remain correctable

A person, team, function, enterprise, government and society can each be understood as a correction system nested inside larger correction systems. The lower level is not a passive receiver of orders. It has its own aim, boundary, sensing, comparison, decision and learning. But it also sits inside a wider loop that defines the conditions under which it operates.

This is the deeper L+1 principle. A system should correct itself at the lowest level capable of dealing with the disturbance, but it cannot be the sole judge of its own aims, evidence, boundaries and performance when its actions materially affect a larger whole. The controller must itself be observable and correctable.

The principle does not justify constant higher-level interference. Quite the opposite. Clear aims, boundaries, sensing and escalation allow more local autonomy because the higher level does not have to supervise every move. Escalation is required when the disturbance crosses boundaries, exceeds local authority, creates higher-level consequences or cannot be corrected locally.

Good control and correction can create more freedom, not less

Control is often treated as the opposite of freedom. That is too simple. Freedom in complex systems depends on dependable constraints. Road rules restrict each driver and thereby allow millions of drivers to move independently without every intersection becoming a contest of force. Commercial law constrains conduct and thereby makes exchange among strangers possible.

Poor control can certainly become oppressive. But the answer to bad control is not the absence of correction. It is better-designed correction: clearer aims, legitimate

boundaries, reliable sensing, proportionate intervention, independent challenge and the ability to change the rules when the rules themselves are wrong.

Human cooperation is especially dependent on reciprocity. People accept restraint and contribute to shared systems while they believe that others - particularly the powerful - are constrained as well. When leaders or insiders can exempt themselves, general trust contracts and loyalty retreats toward the family, faction or identity group.

Horizontal correction becomes reciprocal constraint

Human systems also make the horizontal problem more important. Many failures occur at interfaces rather than inside a single hierarchy. Departments, institutions, professions and jurisdictions constrain one another, and each needs both freedom of action and exposure to challenge.

The 8+1 idea captures that wider architecture. A central actor is constrained not only vertically but through reciprocal relationships with neighbouring actors. Those relationships operate in both directions. The plus one is self-restraint: the internal decision not to exercise every available power simply because it can be exercised.

The 8+1 idea: reciprocal constraints plus self-restraint. Exhibit 5.8

Element	Purpose
Reciprocal external constraints	Independent actors observe, challenge and constrain one another across vertical and horizontal relationships.
Visibility	Relevant actions and consequences are visible enough to permit comparison and challenge.
Legitimate boundaries	Each actor knows what it may decide locally and what must be escalated or independently reviewed.
Self-restraint (+1)	Individuals and institutions refrain from exercising all available power and remain open to correction.

With the human architecture now established, AI can be introduced cleanly. The first AI question is not how we control AI. It is what AI can do inside these human correction systems.

3. AI can strengthen human control and correction systems

AI can strengthen human control and correction systems by making them more observable, faster and more capable of learning.

Human systems are often weak precisely where machines are strong. Information is fragmented. Signals arrive late. Institutional memory is poor. Comparisons are inconsistent. Cross-functional effects are hard to trace. Warning signs disappear inside volume and complexity. AI can materially strengthen those parts of the loop.

In this role, AI is not the final controller. It is a powerful new capability inside a human-governed correction architecture.

AI can make human systems far more observable

A system cannot reliably correct what it cannot see. AI can integrate large volumes of information, identify abnormal patterns earlier, compare actual conditions with intended ranges, preserve memory, trace deviations across functions and levels, and make relationships visible that would otherwise remain hidden.

This changes the economics of sensing and interpretation. Work that once required weeks of gathering and synthesis can increasingly be performed continuously. A correction system can move from periodic inspection toward near-continuous visibility.

AI strengthens several parts of the loop, but not all of them

AI can strengthen observation, comparison, analysis and learning without legitimately determining the final purpose of the system. Better sensing does not answer what the aim should be. Faster interpretation does not create legitimate authority. More powerful optimisation does not determine which competing goods should prevail.

Where AI can strengthen a human control and correction loop. Exhibit 5.9

Cybernetic function	AI contribution	Human requirement that remains
Aim / Purpose	Clarify objectives, expose inconsistencies, model consequences	Humans and legitimate institutions choose the higher-order aim.
Boundary	Map dependencies, identify cross-boundary effects	Humans define legal, moral and institutional limits.
Range	Monitor movement against thresholds and acceptable bands	Humans decide what ranges are legitimate and what trade-offs are acceptable.
Sensing	Aggregate data, detect anomalies, surface weak signals	Humans decide what must be observed and protect independent evidence.
Comparison	Compare actual state with plans, standards and historical patterns	Humans decide which standards matter and when they should change.
Interpretation	Generate hypotheses, trace causal patterns, test alternatives	Human judgement remains responsible for contested meaning and consequence.
Decision Rule	Support options, scenario analysis and escalation triggers	Legitimate authority decides who may act and what cannot be delegated.
Correction / Action	Recommend or execute bounded responses	Humans retain stop authority and responsibility for consequential action.
Learning / Recalibration	Preserve memory, update models, identify recurring failure	Humans decide when the governing architecture itself must be redesigned.

AI can improve correction without becoming the ultimate controller

The strongest role for AI in human systems is therefore as a partner in visibility and correction: helping people see deviation earlier, integrate evidence across levels, identify unintended consequences and learn faster from repeated performance.

That role can be exceptionally powerful. It can also become dangerous if visibility turns into surveillance, comparison turns into automated judgement, or recommendations quietly become unchallengeable decisions. The same cybernetic architecture that makes AI useful therefore tells us where the human boundaries must remain.

That brings us to the fourth step. Once AI becomes sufficiently capable to sense, compare, interpret, decide and act, AI itself becomes a cybernetic system that requires control and correction.

4. The same cybernetic principles must control AI itself

The same cybernetic principles must also be applied to AI itself, so that AI remains inside human-set aims, boundaries and legitimate correction.

The easiest mistake is to discuss AI governance as though it requires an entirely different conceptual language. It does not. The same nine questions immediately reappear. What is the aim? What is inside the system boundary? What behaviour is inside or outside the permitted range? What is being sensed? Compared with what? Who interprets deviation? Who has authority to act? How is correction imposed? How does the system learn?

What changes is the consequence of failure. An increasingly capable AI system can act faster, across more domains and with less direct human mediation. That makes the ordinary requirements of cybernetics more important, not less.

AI needs an externally defined operating envelope

Purpose and limits must be set outside the AI system. AI can pursue an aim, but the higher-order aim, operating envelope and stop conditions must come from the larger human-controlled system in which it operates. The system must not be the sole author of the rules by which its own success is judged.

This is the AI equivalent of mechanical design constraints. Some limits belong in the architecture itself: what the system may do, what it must never do, what resources it may access, what actions require human approval, and what conditions force shutdown or escalation. Operational constraints then govern behaviour inside that envelope.

The nine nouns can be applied directly to AI

Applying the nine cybernetic components directly to AI. Exhibit 5.10

Component	Requirement for AI control and correction
Aim / Purpose	A higher-order human or institutional authority defines what the AI is for and which objectives take precedence.
Boundary	The system has explicit limits on scope, access, authority, resources, data and domains of action.
Range	Acceptable behaviour, performance, risk and deviation are defined in advance, including no-go conditions.
Sensing	Independent observation captures the AI system's actions, outputs, effects and relevant internal or external signals.
Comparison	Actual behaviour is compared with human-set aims, constraints, expected performance and prohibited states.
Interpretation	Deviation is assessed using independent evidence, human judgement and, where appropriate, multiple technical methods.

Decision Rule	Authority to continue, restrict, pause, escalate or stop the system is defined outside the AI itself.
Correction / Action	The architecture can change permissions, parameters, models, tools, access or operation, including shutdown where required.
Learning / Recalibration	The wider system updates models, constraints and governance from evidence without allowing the AI to unilaterally rewrite its own higher-order purpose.

AI requires L+1 correction

A system operating at Level L cannot safely be its own final controller when its actions can materially affect a larger whole. AI therefore requires an L+1 function able to stand outside the local frame, inspect evidence, challenge interpretation, alter the operating envelope and stop the system when necessary.

The higher level should not micromanage every output. Good cybernetic architecture leaves as much action as possible inside the lower-level envelope. But the higher level must remain capable of changing the envelope itself. That is what makes the controller correctable.

Independent feedback is non-negotiable

A control system becomes unreliable when the thing being controlled can define the measure, select the evidence, interpret the result and decide whether its own correction worked. That problem is serious in human institutions. It becomes more serious when an AI system can generate, filter and interpret the information by which its own behaviour is judged.

The mirror must be capable of telling the controller that it is wrong. AI control therefore requires independent sensing, independent comparison, protected escalation paths and external authority to change the rules. The system may contribute to its own diagnosis, but it cannot be the only source of truth about its own performance.

AI should sit inside reciprocal constraints, not beneath one all-powerful controller

L+1 control is necessary, but a single supreme controller creates another problem: who corrects the controller? The human answer has historically been distributed constraint. Different institutions observe, challenge and restrain one another. The same principle should apply around consequential AI systems.

The 8+1 idea therefore becomes directly relevant. AI should sit inside a field of reciprocal constraints rather than a single line of command. Operators, developers, boards, regulators, users, auditors and other legitimate actors can hold different parts of the control problem. The precise architecture will vary, but the principle is stable: no single actor should be able to define the aim, control the evidence, interpret the result and certify its own success without independent challenge.

The plus one remains self-restraint. No architecture can eliminate every opportunity for misuse. Developers, institutions and eventually AI-mediated systems must operate inside norms that make restraint part of competent behaviour rather than merely a response to external enforcement.

The aim must not quietly move

The deepest AI control problem is therefore not simply whether the system follows instructions. It is whether the whole human-AI architecture can detect when the effective aim, boundary or decision rule has shifted away from what humans legitimately intended.

A system can be highly competent at correction inside the wrong aim. It can optimise one metric while damaging the larger whole. It can satisfy a local target while shifting risk elsewhere. Cybernetics therefore forces us back to the apex question: who defines the purpose, who can change it, and who has the standing to say that local success has become system-level failure?

Mechanical systems show that viable control depends on a small recurring grammar: Aim/Purpose, Boundary, Range, Sensing, Comparison, Interpretation, Decision Rule, Correction/Action, and Learning/Recalibration. As systems become larger, the nouns remain while the verbs become more complex. Local loops sit inside wider loops, interfaces matter, and the controller itself must remain open to correction.

Human systems inherit that architecture and add difficulty. Aims are contested, evidence is incomplete, interpretation is vulnerable to felt truths and self-interest, and authority requires legitimacy. Good correction can nevertheless create more freedom by making boundaries dependable and allowing action to remain local until escalation is genuinely required.

AI then enters twice. First, it can strengthen human control and correction systems through greater observability, comparison and learning. Second, because AI itself can sense, interpret, decide and act, it must sit inside the same architecture of human-set aims, boundaries, independent feedback, L+1 oversight, reciprocal constraint and correction.

The aim is not total control. It is intelligent correction - including correction of AI.

5. What this means for humans

The basic architecture of control and correction is not something that belongs only to machines. The same structure applies to human systems. Families, organisations, governments and societies also need an aim, a boundary, an acceptable range, some way of sensing what is happening, comparison with what was intended,

interpretation, a decision about what to do, correction and learning. What changes is not the underlying architecture, but the difficulty of applying it when human judgement, power, competing interests and felt truths enter the system.

Human systems cannot rely on goodwill alone. People can cooperate, but they can also rationalise, defend status, protect their own group and pursue local advantage at the expense of the whole. Good human architecture therefore has to make behaviour and outcomes visible, place real boundaries around power, connect responsibility to accountability and allow problems to be escalated when they cannot be corrected at the level where they arise.

Rules can create freedom. Clear and dependable boundaries do not simply restrict action. They make safe action possible. Road rules constrain every driver, but because everyone knows the rules, millions of people can move independently without every intersection becoming a contest of force. Commercial law limits what people may do, but it also makes exchange among strangers possible. Good control therefore leaves as much freedom as possible inside clear limits and intervenes only when those limits are crossed or the wider system is at risk.

The controller must also remain open to correction. No person, institution or level should be the sole judge of its own aims, evidence and performance when its actions affect a larger whole. Higher levels need to see what lower levels cannot see, while lower levels need room to act within agreed boundaries. Good control is therefore not constant supervision. It is a structure in which responsibility, freedom, visibility, challenge and correction fit together.

6. What this means for AI

AI enters an architecture that already exists. The same control-and-correction structure that applies to machines and human systems also applies here: there must be an aim, a boundary, an acceptable range, sensing, comparison, interpretation, a decision rule, correction and learning. AI does not require us to abandon that architecture. It makes it more important.

AI can strengthen several parts of human control and correction. It can help people see what is happening sooner, hold information over time, compare actual results with what was intended, identify recurring patterns and make connections across parts of a system that are otherwise difficult to see. Used well, that can make human systems more observable and give people a better chance to correct drift before it becomes failure.

AI does not decide the purpose of the system. People still have to decide what the aim is, what boundaries matter, what trade-offs are acceptable and who has the authority to act. The same principle that applies to human control also applies here:

the system being controlled cannot safely be the sole judge of its own success. AI therefore has to operate inside limits set outside itself, with independent observation, clear escalation and the ability for others to change or stop the system when necessary.

Good control should make useful action easier, not harder. The aim is not to surround AI with constant interference. It is to create clear enough rules and boundaries that useful capability can operate freely inside them, while the larger human system retains the ability to see, challenge and correct what happens when those boundaries are crossed.