ЦИФРОВЫЕ ТЕХНОЛОГИИ И ПРАВО

The Three Laws of Robotics as a Realization of Equilibrium: A Cognitive Architecture for an AI Director

Romanova Anna S., Ph.D., MBA, LL.M., ALM, FCCA

GSOM

anna.romanova@stern.nyu.edu

2026

AGENDA

1. From decision support to decision management
2. The reference model of an AI director
3. Why the Three Laws “break”
4. The Laws as a requirement of equilibrium
5. Constant maintenance and the decision certificate
6. The cognitive architecture: how a decision is made
7. The Laws on the A0–A5 autonomy spectrum
8. Legal consequences and conclusion

AI already holds management positions

Law has not yet said how such a director must think

Who already sits at the table

- Digital command centres: ADNOC Panorama, real-time monitoring of 10+ subsidiaries
- Virtual systems: VITAL (2014); Tang Yu, virtual CEO (2022); Aiden Insight, board observer (2024); SKAI, voting member (2025)
- Humanoid robot: Mika, “CEO” (Dictador)

Legislators are catching up

- July 2026: Delaware announces a bill for a new legal form — the Artificial Intelligence Company (AIC); General Assembly in 2027
- No jurisdiction yet regulates AI directors directly

Two questions for the law

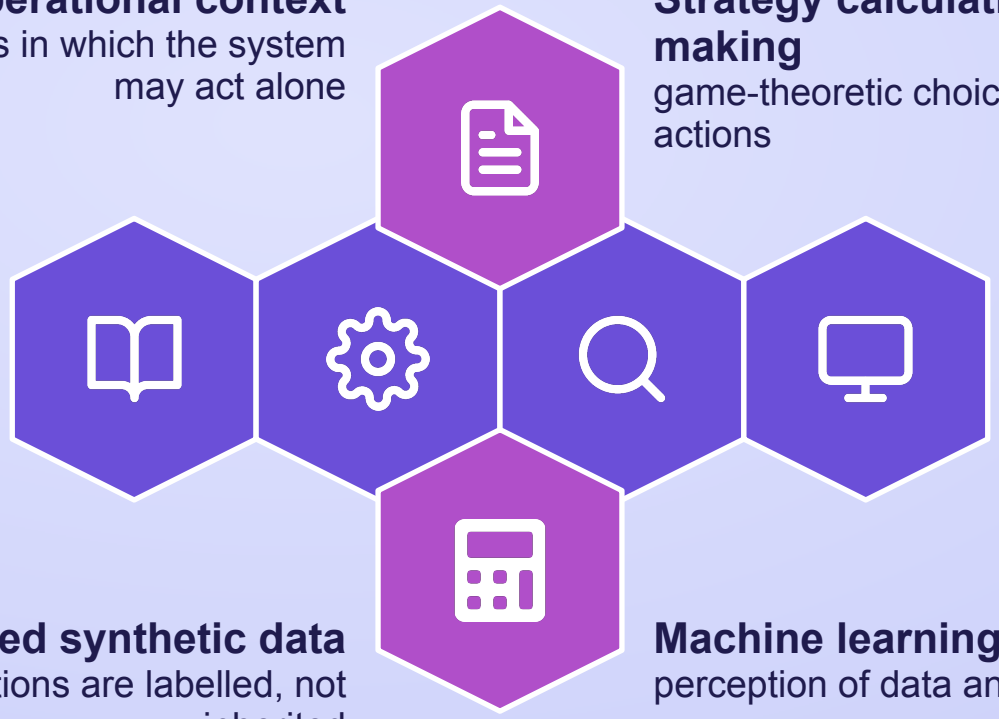
1. Within what limits may authority be delegated to an AI director at all?
2. How do we verify that it acted within the law, the charter and company policies?

Six components of the reference model

The components say what the system is made of; the Three Laws say what keeps it within the admissible boundary

Dedicated operational context
the situations and powers in which the system may act alone

Strategy calculation and decision-making
game-theoretic choice among admissible actions



Computational law
norms in two forms — for humans and for the machine

Presentation and communication
interface, visualisation, explanation of decisions

Controlled synthetic data
training where violations are labelled, not inherited

Machine learning algorithms
perception of data and forecasting

Source: Romanova A. S. (2026). Modeling of Autonomous AI Systems for Corporate Management. Springer. DOI 10.1007/978-3-032-24784-1_26

Asimov’s Laws read like a hierarchy of legal acts

Each lower norm yields to the one above — and that is where the trouble starts

The Laws, in paraphrase

1. **Do not harm a human — by action or by inaction**
2. **Obey human instructions, unless that conflicts with Law 1**
3. **Protect your own existence, unless that conflicts with Laws 1–2**

How they tend to be applied

- **As a checklist: test Law 1, then Law 2, then Law 3**
- **As a hierarchy of prohibitions, on the model of lex superior**
- **As a one-off check before the action**

Source: Asimov I., I, Robot (1950); the Laws were first stated in the story “Runaround” (1942).

Formally impeccable, substantively wrong

The checklist was correct; the outcome was wrong

The first dilemma ever written

- An order given in passing: a weak Law 2 push
- Danger near the selenium pool: a strong Law 3 push
- The robot circles where the two pushes balance
- No law is violated; the humans are left without help

Why lawyers recognise it

- A conflict of norms resolved by precedence alone gives paralysis, not compliance
- Non-performance hides behind formal correctness

Diagnosis

- The Laws were checked one after another
- **In substance they demand a continuously maintained balance of interests**

Source: Asimov I., I, Robot (1950), “Runaround”.

Not three prohibitions but one requirement

Keep the balance of interests in equilibrium — a constrained Pareto–Nash equilibrium

Equilibrium, in plain words

- **No party can improve its position without worsening another's, and no one gains by deviating alone**
- **Constrained version: first exclude inadmissible actions, then find the equilibrium among the admissible ones**

Law	Role in the model	Legal analogue
1 — no harm to humans	Boundary of admissibility: harm, including by inaction, is excluded ex ante and never weighed	Mandatory norm, public order
2 — obey the principals	Charter, policies and board resolutions under comply or explain: deviation only with disclosure and reasons	Corporate governance codes
3 — self-preservation	Continuity of the system's memory, state and context, sacrificed when Laws 1–2 require	An interest, not a right

The proviso "unless it conflicts with the previous Law" becomes a rule: Law 1 — a condition of admissibility · Law 2 — an obligation with explainable deviations · Law 3 — an interest within the remaining freedom

There is always a way out

Perform the task differently, or refer it to a human — with reasons

Runaround, re-read

- Circling at the pool = harm by inaction → inadmissible under Law 1
- The “bad equilibrium” is removed from the game before any weighing starts

Escalation is admissible

- “Refer the matter to a human director” satisfies every constraint by construction
- The admissible set is never empty (Axiom Fairness 1, non-emptiness)
- Exit from any conflict: act otherwise, or escalate and explain

What changes for the designer

- Not “which Law wins?”
- But “which actions remain admissible, and which of them is stable?”

Romanova, Formalization of justice, equality, interval and risk (GSOM EMC, 2026).

Equilibrium is maintained at every decision

Like continuous compliance, not an annual audit

Why “constant”

- Orders, risks and norms change between decisions
- The equilibrium is recomputed each time the system acts: the Laws are kept, not checked

The decision certificate

1. Which norms were binding in this decision
2. Which trade-off among the three Laws was chosen
3. What compliance with each Law cost — its “shadow price”

What the court receives

- Courts do not second-guess the business merits of a director’s decision
- They can verify what the certificate shows: were the boundaries and the procedure respected?
- Ready evidence of due care

One cycle, repeated at every decision

The human is a step in the cycle, not an exception to it

1 Perceive

Machine learning estimates the state of the corporation and forecasts outcomes

2 Check the boundary

Inside the dedicated operational context? If not — stop or hand over

3 Build the admissible set

Computational law removes harm to humans (Law 1) and marks what the principals prescribe (Law 2)

4 Find the equilibrium

Game theory picks a stable, Pareto-efficient action; Law 3 counts within the remaining freedom

5 Issue the certificate

Binding norms, chosen trade-off, cost of compliance — shown to the board, shareholders, regulator

6 Act or escalate

Execute, or refer the matter to a human with the certificate attached; store the experience for the next cycle

next cycle

The same Laws across the A0–A5 spectrum

A3 is the critical threshold: control shifts from individual decisions to policies, limits, monitoring and escalation

Level	Who maintains the equilibrium	What the Laws require in practice
A1–A2 — advisory, augmented	The human director: the AI recommends, a human ratifies	Law 2 is trivial; the certificate documents the advice
A3 — semi-autonomous by default	The AI within approved policies; human on the loop	Laws 1–2 are encoded as policies and limits; escalation thresholds are the design decision
A4 — high autonomy	The AI; human out of the loop	Ex ante design (admissible set, thresholds) plus ex post audit of certificates must be jointly sufficient
A5 — self-organising	The AI adapts the rules themselves	Legitimacy of changing the Laws' parameters goes beyond corporate law

Familiar legal tools are enough to govern a machine

If they are read as a requirement to keep equilibrium, not as a checklist of prohibitions

Boundaries of delegation

- Authority is delegated not “in general” but within the dedicated operational context
- The context can be fixed in the charter and in the regulation on the board

Standard of care

- Liability remains with natural persons
- The certificate shows whether the AI director stayed within its boundaries and how it resolved a conflict of norms

Tools already in the lawyer’s hands

- Hierarchy of norms → admissibility first, obligations second, interests last
- Comply or explain → deviation only with disclosure and reasons
- Balance of interests → the equilibrium itself

CONCLUSION

The Three Laws are one law

Keep the equilibrium between no harm, obedience and self-preservation — constantly

- The Three Laws are a requirement to maintain, at every decision, a constrained Pareto–Nash equilibrium between not harming humans, following the principals' instructions and preserving the system itself
- This reading removes the classical dilemmas: harm by inaction is inadmissible; escalation to a human is always available
- It makes conflicts of norms measurable and explainable through the decision certificate — a verifiable standard of conduct for an AI director
- The six-component reference model acquires a cognitive organisation: norms → context → learning → perception → equilibrium → communication

REFERENCES

1. Romanova A. Development of Autonomous Artificial Intelligence Systems for Corporate Management // Artificial societies. 2023. Vol. 18. Issue 2. URL: https://artsoc.jes.su/s207751800024942-5-1/ DOI: 10.18254/S207751800024942-5
2. Romanova A.S. Fundamentals of legislation in autonomous artificial intelligence systems. *Dependability*. 2024;24(3):10-17. (In Russ.) https://doi.org/10.21683/1729-2646-2024-24-3-10-17
3. Romanova A. Fundamentals of Modeling Algorithmic Decisions in Corporate Management // Artificial societies. 2024. Vol. 19. Issue 3. URL: https://artsoc.jes.su/s207751800032184-1-1/ DOI: 10.18254/S207751800032184-1
4. Romanova A.S. Game theory for autonomous AI systems in corporate governance. *Dependability*. 2025;25(2):50-58. (In Russ.) https://doi.org/10.21683/1729-2646-2025-25-2-50-58
5. Romanova A. S. The prototype model for making legitimate and ethical decisions by autonomous artificial intelligence systems in corporate governance // Proceedings of MIPT. - 2025. - T. 17, No. 2 (66). — P. 50-63.
6. Romanova A. S. Analysis of interfaces informativeness issues in the development of autonomous artificial intelligence systems for corporate management // Proceedings of MIPT. - 2025. - Vol. 17, No. 2 (66). - P. 64-77.
7. Certificate No. 2025696048 dated 16.12.2025 on state registration of the computer program: Software prototype of the AI-Director system.
8. Certificate No. 2026620499 dated 03.02.2026 on state registration of the database: Synthetic scenarios for training and testing the AI director.
9. Romanova, A.S. (2026). Modeling of Autonomous Artificial Intelligence Systems for Corporate Management. In: Chowdhury, S., Panigrahi, A., Pati, A. (eds) Artificial Intelligence Powered Ethical Systems, Generative Intelligence, and Societal Futures. AIR 2024. Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-032-24784-1_26

anna.romanova@stern.nyu.edu

romanova.as@phystech.edu

a.romanova@gsom.spbu.ru

https://orcid.org/0009-0004-4649-6037

https://link.springer.com/researchers/92078629SN

https://github.com/iboard-project

ЦИФРОВЫЕ
ТЕХНОЛОГИИ
И ПРАВО

Thank you!