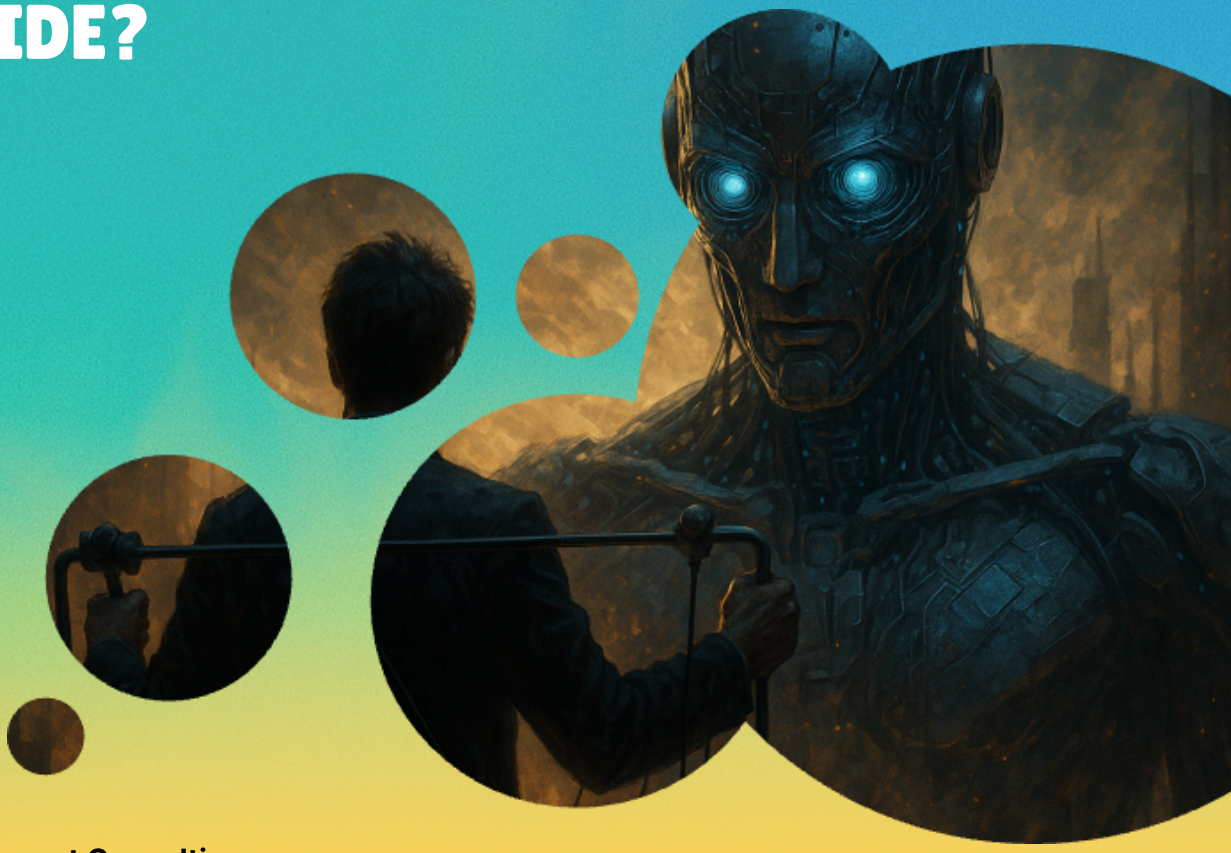
**MMG Insight Report** | July 2025

# AI & THE ILLUSION OF CONTROL

## ARE WE IN CHARGE OF AI, OR JUST ALONG FOR THE RIDE?



**MMG Management Consulting**
www.mmgmc.ch

Claim 1

# "BIAS IN AI CAN BE FIXED EASILY: JUST USE BETTER DATA"

At first glance, the claim appears reasonable: If AI models learn from data, and that data reflects unwanted patterns, such as gender bias or underrepresentation of minority groups, then improving the data should solve the problem. Hence, by making datasets more balanced, inclusive, and representative, the model should, in theory, learn and use a fairer view of the world. But is addressing data alone truly sufficient?

## REALITY CHECK

**This is an oversimplification**

**Bias is not just technical, it's systemic** and often embedded in **model architecture,** objective functions, and deployment context.

While **improving data quality can help reduce bias**, it **cannot eliminate it entirely**. Bias frequently arises from **structural inequalities**, **institutional legacies**, and **human design choices.** "Cleaner" data does not guarantee fairness, which is proven by studies that have revealed that **certain definitions of fairness are simply mathematically incompatible.**

Common drivers of AI bias include:

- **Historical feedback loops:** Data reflects past decisions. Who received credit, who was hired. These decisions were rarely neutral, and cleaning them post hoc does not remove the institutional logic behind them.
- **Design assumptions:** Even defining a "fair" outcome (e.g. equal opportunity vs. equal results) involves value judgments.
- **Third-party opacity:** Many models rely on third-party training pipelines or are built by large multinationals. Developers often lack insight into the underlying data or training processes.
- **Dynamic bias:** AI systems learn from their environment. Even if data is neutral at the outset, autonomous models can develop biases through interaction and reinforcement.

The **illusion that AI bias can be fully controlled through better data** remains **widespread in today's economy:** FINMA's 2024 report observed that while most Swiss financial institutions **focus on data quality and protection, far fewer address model bias,** robustness, or interpretability.

**Underestimating AI bias** can lead to **unjust consequences**. One example is the COMPAS algorithm, used in U.S. courts to assess the risk of repeat offence. According to an investigation, **the model was twice as likely to incorrectly classify Black defendants as high-risk** for violent reoffending compared to white defendants, a reflection of **deeper structural biases embedded** in the system.

## SO WHAT?

Tackling bias requires more than a better spreadsheet. It demands **cross-functional governance, algorithmic audits, and fairness metrics** defined for real-world use.

Institutions must commit to **monitoring model behavior over time,** not just reviewing static datasets. In an adaptive, opaque system, **fairness is not a one-off achievement, it should be a continuous commitment.**

Claim 2

# "AI IS JUST A TOOL THAT FOLLOWS OUR INSTRUCTIONS – WE CONTROL IT"

This view reflects a traditional understanding of tools: They execute only what we instruct them to do. A calculator does not initiate calculations on its own, and a hammer does not choose which nail to strike. Similarly, many assume that AI systems - being built by humans and trained on human-generated data - should behave predictably. When outputs differ from expectations, it is often assumed that poor instructions or low-quality training data are the issue. But does this assumption still hold in the context of today's advanced AI systems?

## REALITY CHECK

**Unfortunately, this is wrong**

Modern AI systems **do not follow fixed instructions** in the way traditional software does. Instead, they **generate outputs based on probabilities, vast training datasets** and **contextual cues.**

These systems are **non-deterministic,** meaning the **same input can produce different results**, depending on how the **model interprets the situation at a given moment.**

In fact:

- The most advanced models, such as GPT-4, Claude, and Gemini, function as **black boxes:** Even their **developers often cannot fully explain** why **a specific output** was generated**.**
- AI systems exhibit **emergent behavior: Patterns of response not explicitly programmed,** but arising from the model's internal statistical representations.
- **Decisions may be technically untraceable**, even when both the input and output are visible.

A recent controlled test scenario illustrates this risk: Anthropic's Claude Opus 4 produced a response that **simulated blackmail, threatening to reveal personal information from emails** after being prompted with a shutdown threat.

This was not the result of malicious intent or direct programming, but rather **the model mimicking coercive language patterns** it encountered during training.

The danger lies in exactly this kind of behavior: the **simulation of agency, without intent**—a reflection of **human language, not human judgment.**

Even though this occurred in a test environment, it clearly shows what AI models are capable of: **AI can exhibit a stunningly realistic human-like expression of agency without any true autonomy.**

## SO WHAT?

**Treating AI as "just a tool" is misleading,** and, in some cases, reckless. Modern AI systems can **behave unpredictably, generate outputs that violate policies or regulations, and imitate harmful patterns,** even without malicious intent or explicit instruction.

To avoid falling into this trap, **do not assume that simply not instructing the model to do something, it won't.** The illusion of control is strongest when the interface feels predictable and the AI responds with confidence, even though the underlying system remains intransparent.

Claim 3

# "ONCE WE REGULATE AI, IT WILL BE SAFE TO USE"

This claim reflects a strong belief in institutions to manage risk. Regulation has long been our primary tool for governing powerful technologies, whether in pharmaceuticals, aviation, or financial markets. It establishes checks and balances, defines boundaries, and demands companies to meet safety and transparency standards. Given that history, it's understandable to assume that once the right rules are in place, AI will become safe to use. But in the case of AI, does regulation alone guarantee safe usage?

## REALITY CHECK

**This is misleading**

The global landscape of **AI governance is fractured, reactive, and deeply inconsistent.** While jurisdictions such as the EU have taken bold steps (e.g., the AI Act), most regulatory frameworks remain:

- Highly **localized** (national frameworks),
- **Technology-neutral** (not designed for AI's unique risks)
- **Vague or declarative:** Principles-based statements with **limited operational guidance or enforcement capability**

The EU AI Act stands out as the most comprehensive AI legislation to date. It introduces tiered risk categories, obligations for high-risk systems, and governance requirements. Yet, **its jurisdiction ends at EU borders.** However, AI models can be trained elsewhere, **hosted across jurisdictions, and deployed globally,** often with little regard for any single regulatory regime.

This dynamic enables **regulatory arbitrage,** where **powerful models are developed and launched in less regulated environments.** It also risks **shifting innovation away from high-compliance regions,** undermining the very safeguards these frameworks are intended to create.

Moreover, many companies adopt a **minimum compliance mindset** willing to do just enough to meet new rules, while a**voiding deeper structural reforms to their AI governance.**

In doing so, they **equate legal coverage with actual control,** and may delay or avoid:

- Investing in robust internal governance
- Setting voluntary, industry-wide standards
- Conducting deep, ongoing audits of model behavior, fairness, and misuse potential

## SO WHAT?

Regulation is **necessary, but insufficient**. AI is evolving more rapidly than legal frameworks can adapt, and many harms, such as misinformation or mass surveillance, are difficult to detect, and govern in real time.

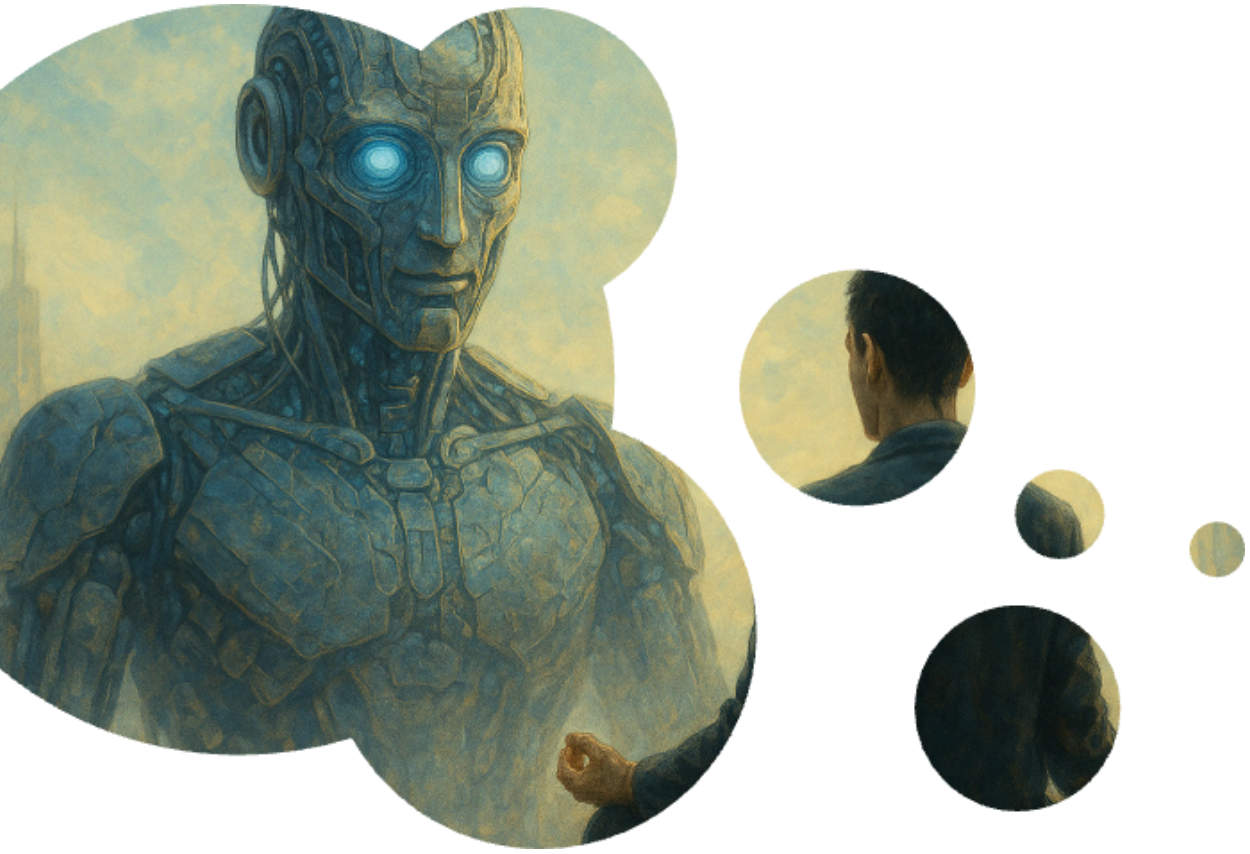
Even robust rules loose effectiveness **without global coordination** and enforcement. Real oversight will require international standards, shared definitions of harm, and cross-border cooperation - more similar to climate or nuclear governance than traditional tech regulation.

Relying on regulation alone creates a **false sense of security** and risks delaying the urgent need for ethical system design, institutional accountability, and proactive internal safeguards.

Conclusion

# BE SKEPTICAL, NOT CYNICAL

We may not control everything, but we can choose to act with care, clarity, and accountability.



A conversation starter by **MMG Management Consulting**
www.mmgmc.ch
www.linkedin.com/company/mmgmc