The Safety-First Rule — AI SAFE 1: Why AI Must Put Human Safety Before Efficiency

By Michal Florek

Executive Summary

Artificial Intelligence promises efficiency, speed, and optimisation. But when efficiency is prioritised over safety, the results can be catastrophic — from the loss of human life in autonomous vehicles to sudden collapses in financial markets.

This paper proposes the **Safety-First Rule**, the first of five AI SAFE frameworks. It argues that all AI systems deployed in critical areas must embed non-bypassable safety mechanisms, undergo independent certification, and face stress testing before deployment.

Two case studies — the **Uber self-driving car fatality (2018)** and the **2010 Flash Crash** — illustrate that efficiency-driven AI without brakes creates harm for individuals and instability for societies. The paper closes with a roadmap matrix of challenges, gaps, and opportunities, and five clear policy recommendations.

Section 1: The Problem of Efficiency Without Safety

AI systems are designed to optimise. They seek the fastest route, the lowest cost, or the highest profit. But efficiency is not the same as safety. Without guardrails, efficiency becomes dangerous: cars crash, markets collapse, and health systems misdiagnose.

Section 2: Case Study 1 — The Uber Self-Driving Car Fatality (2018)

In March 2018, a self-driving car operated by Uber struck and killed pedestrian Elaine Herzberg in Tempe, Arizona.

- The car's sensors misclassified her path multiple times.
- Emergency braking had been disabled for smoother performance.
- The safety driver was distracted, with only 1.3 seconds to react.

Lesson: Convenience (smooth rides) was prioritised over safety (emergency brakes). The lack of independent oversight for auto-pilot logics design had allowed this decision to proceed unchecked.

Section 3: Case Study 2 — The 2010 Flash Crash

On 6 May 2010, global financial markets plunged within minutes, wiping out nearly $1 trillion in value before rapidly rebounding.

- Algorithmic trading systems interacted in ways their designers had not foreseen.
- No central safety mechanism or "circuit breaker" was in place.
- Market stability was sacrificed for trading speed and efficiency.

Lesson: Systemic risk arises when efficiency-driven algorithms operate without brakes, magnifying instability at scale. With risk impact increasing where algorithm used has been programmed by ill motivated individuals.

Section 4: Bridging Argument — From Human Harm to Systemic Harm

At first glance, a fatal crash in Arizona and a global financial meltdown seem unrelated. But both reveal the same pattern: **when AI pursues efficiency without brakes, harm follows.**

- In Tempe, a life was lost.
- On Wall Street, global trust in markets was shaken.

Efficiency does not ask: *Is this safe?* It only asks: *Is this faster?*
This is why AI must embed safety as the first principle.

Section 5: Challenges & Gaps Matrix – Framework 1: Safety-First Rule

Domain	Current Challenge	Key Gaps	Opportunities	Overlaps	Legislation / References
Autonomous Transport	Safety mechanisms disabled for performance (Uber case).	No global standard for non-bypassable overrides.	Build ISO-style certification for "AI Safety Locks."	Transparency & Audit (F2).	EU AI Act (Art. 9), NHTSA AV Guidelines.
Financial Systems	Algorithms optimise for speed, ignore stability (Flash Crash).	Weak requirements for stress testing.	Adapt Basel-style stress tests for algorithmic trading.	Systemic Resilience (F4).	Basel III, MiFID II, FCA controls.
Healthcare AI	AI makes treatment calls without human fail-safes.	No mandatory "human override" protocols.	Create digital health kill switches.	Human Oversight (F3).	FDA AI/ML Device Action Plan (2021).
Public Safety AI	Predictive policing prioritises efficiency over fairness.	No universal bias/safety validation pre-deployment.	Independent audits pre-rollout.	Transparency (F2), Ethics (F5).	EU AI Act (High-Risk AI).
Infrastructure / Utilities	AI optimises for cost, ignores cascading failures.	Missing "fail operational" design.	Embed safety-first in national resilience standards.	Systemic Resilience (F4).	UK NIS Regs.
Cross-Domain Governance	Companies self-certify AI safety.	No harmonisation of certification.	International "AI Safety Mark."	Transparency (F2), Ethics (F5).	OECD AI Principles, G7 Hiroshima Process.

Insights:

- Safety overrides are inconsistently enforced.
- Certification is fragmented across sectors.
- Human oversight is often impractical by design.
- Stress testing exists in finance but not in other high-risk areas.
- AI needs to be shielded from being abused by bad actors.

Section 6: Policy Recommendations

1. Non-Bypassable Safety Mechanisms

Emergency brakes and safety overrides must be mandatory.

🔗 EU AI Act (2024), Article 9.

2. Independent Certification Before Deployment

Critical AI must be tested and certified by third parties.

🔗 UK AI Safety Summit framework (2023).

3. Human Oversight With Realistic Design

Humans need adequate time and alerts to act.

🔗 NHTSA AV Policy (2017).

4. Transparency and Traceability

Critical AI must log and explain decisions for audits.

🔗 OECD AI Principles (2019).

5. Mandatory Stress Testing

Scenario-based resilience testing across sectors.

🔗 Basel Committee on Banking Supervision.

Conclusion

The first rule of AI governance must be safety. Efficiency without brakes is not progress — it is recklessness. By embedding the Safety-First Rule into law, certification, and design, societies can prevent harm, protect markets, and build trust in AI as a tool for collective good.
