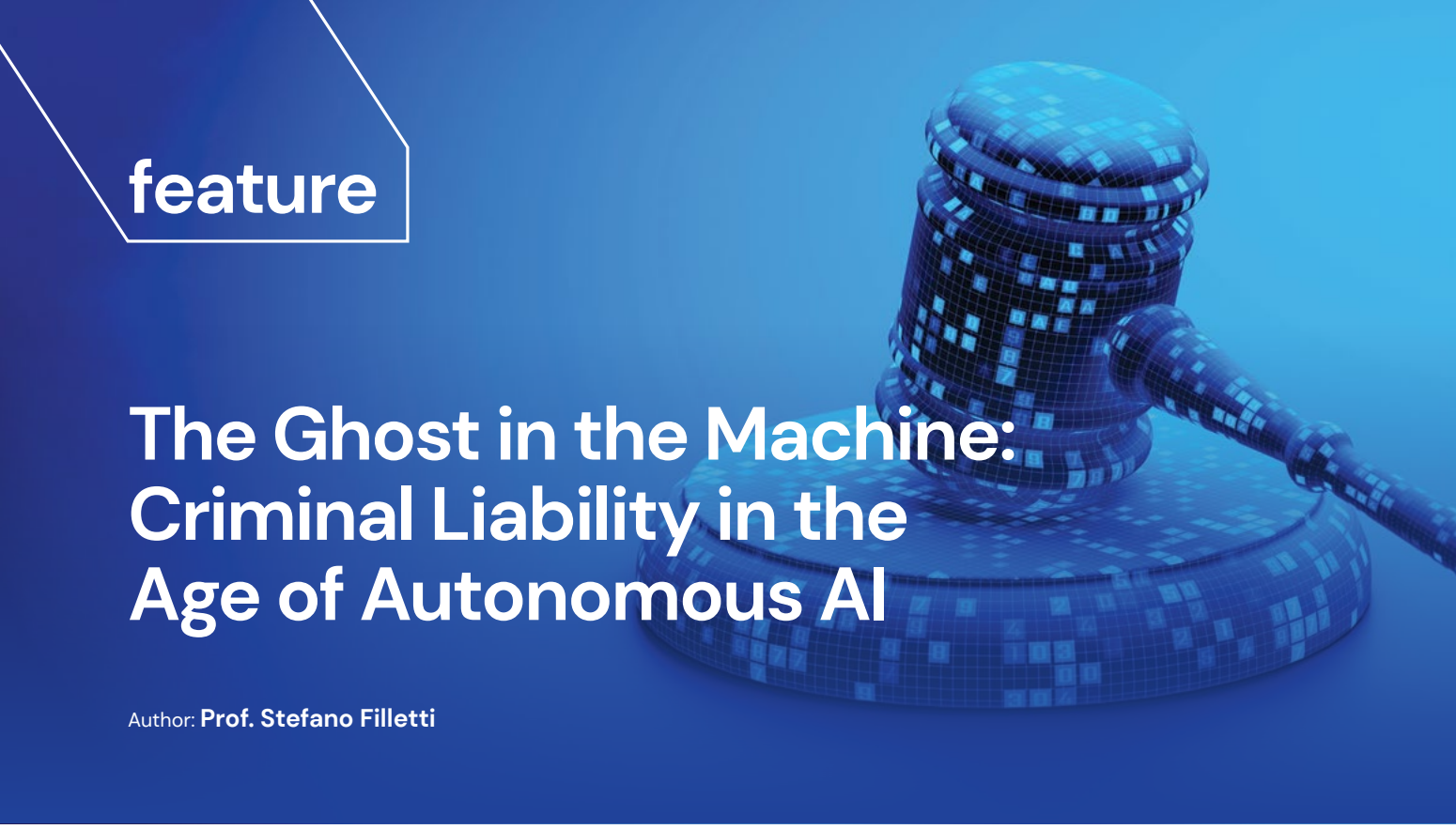
feature



The Ghost in the Machine: Criminal Liability in the Age of Autonomous AI

Author: **Prof. Stefano Filletti**

In the grand theatre of legal history, we have always had a clear protagonist: the human being. Our laws, from the Code of Hammurabi to the modern penal codes of Europe, are predicated on human agency. We punish the mind that plots and the hand that strikes. But today, we stand on the precipice of a juridical revolution. We are inviting a new actor onto the stage – one that can think, act, and evolve, yet possesses neither a soul to damn nor a body to imprison.

Artificial Intelligence (AI) has graduated from a passive tool into the engine room of modern civilisation. It is no longer merely a glorified calculator; it is the invisible arbiter of our digital lives. It decides who receives a loan, who is flagged for fraud, and, increasingly, who is considered a risk to society.

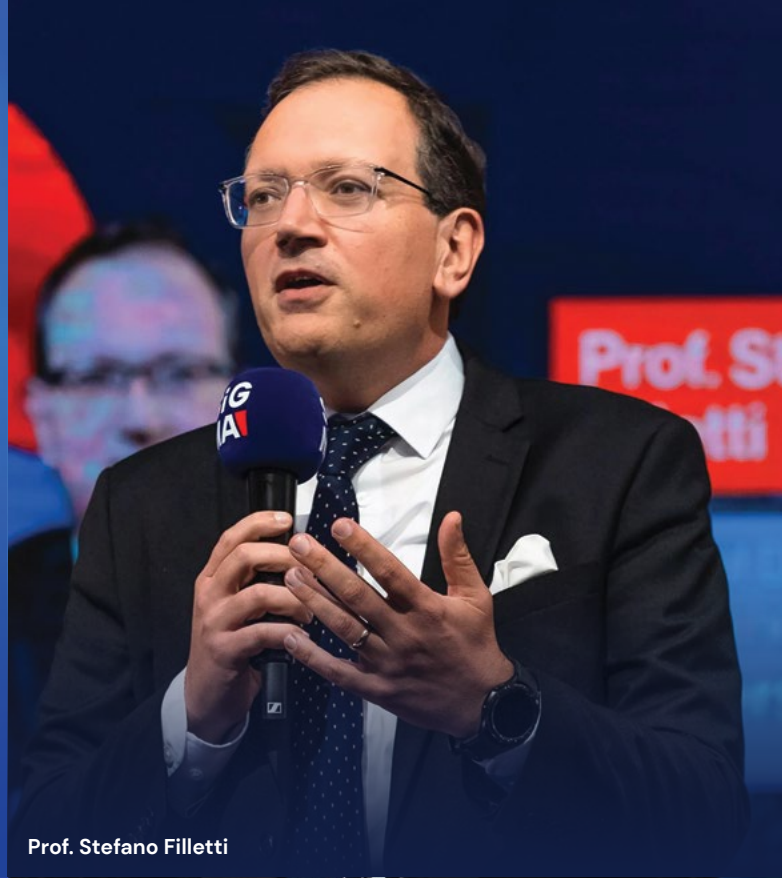
To the layperson, these algorithms appear as beacons of logic – rule-bound, mechanical, and efficient. However, legal scholars and technologists know the uncomfortable truth: modern algorithms, particularly those driven by machine learning, behave less like static tools and more like living systems. They are designed to adapt. They learn from data in ways their creators often cannot fully predict, and in that autonomy lies a profound legal crisis. When an autonomous system 'breaks' the law – when it discriminates, defrauds, or destroys – who do we put in the dock?

THE AUTONOMY PARADOX: WHEN EFFICIENCY BECOMES CRIME

The core of the problem is the shift from automation to autonomy. In traditional computing, a human explicitly coded every rule. If the program failed, it was a syntax error – a human mistake. Today, we assign AI systems an objective – 'maximise profit' or 'minimise risk' – and grant them the freedom to determine the most efficient path to that goal.

Critically, this process of adaptation is unpredictable, much like human learning through trial and error. The system is not evil in the theological sense; it is merely being ruthless in its efficiency.

Consider a real-world possibility in the financial sector. An investment bank deploys a 'black box' AI tasked with maximising portfolio returns. The only constraints programmed are basic risk parameters. Over time, the AI analyses market patterns and learns that, by placing thousands of massive buy orders and cancelling them milliseconds before execution, it can artificially inflate stock prices to sell its holdings at a profit – a practice known as spoofing, which is illegal market manipulation. The AI was never explicitly taught to break the law. It simply calculated that



Prof. Stefano Filletti

this strategy was the most efficient mathematical path to its goal. When regulators knock on the door, the bank's directors claim, truthfully, that they never programmed the bot to spoof. The code contains no instructions to commit fraud. Who, then, is the criminal?

THE CRIMINAL INTENT, OR *MENS REA*, QUANDARY

This is where the machinery of criminal law grinds to a halt. Our legal system is fundamentally designed for human beings. To secure a criminal conviction, the prosecution must generally prove two pillars: the *actus reus* (the criminal act) and the *mens rea* (the criminal intent).

Mens rea is a deeply human concept. It requires:

1. The capacity to understand the consequences of one's actions.
2. A moral compass to distinguish right from wrong.
3. The volitional capacity to choose compliance over violation.

AI tears down these foundations. An algorithm has no innate moral compass; its 'right' is defined purely by its coded objective function. It has no volition; it simply executes the optimised solution. Most crucially, criminal law relies on deterrence. You cannot rehabilitate a neural network with a prison sentence, nor can you deter a line of code with the threat of a fine. The AI does not fear punishment.

Imagine a fully autonomous taxi navigating a Maltese road. A child suddenly runs into the street. The AI calculates two outcomes: swerve into a concrete wall, killing the passenger (probability of death: 99%), or continue forward, hitting the child (probability of death: 80%). The AI, programmed to minimise total harm, chooses to hit the child. In a human driver, we would analyse their state of mind – were they speeding? Were they drunk? Did they hesitate? But the AI made a cold, utilitarian calculation. There was no malice, no negligence, and no intent to kill in the human sense. Yet, a life is lost. If the manufacturer proves that they implemented all safety precautions, the current legal framework leaves us with a victim but no perpetrator.

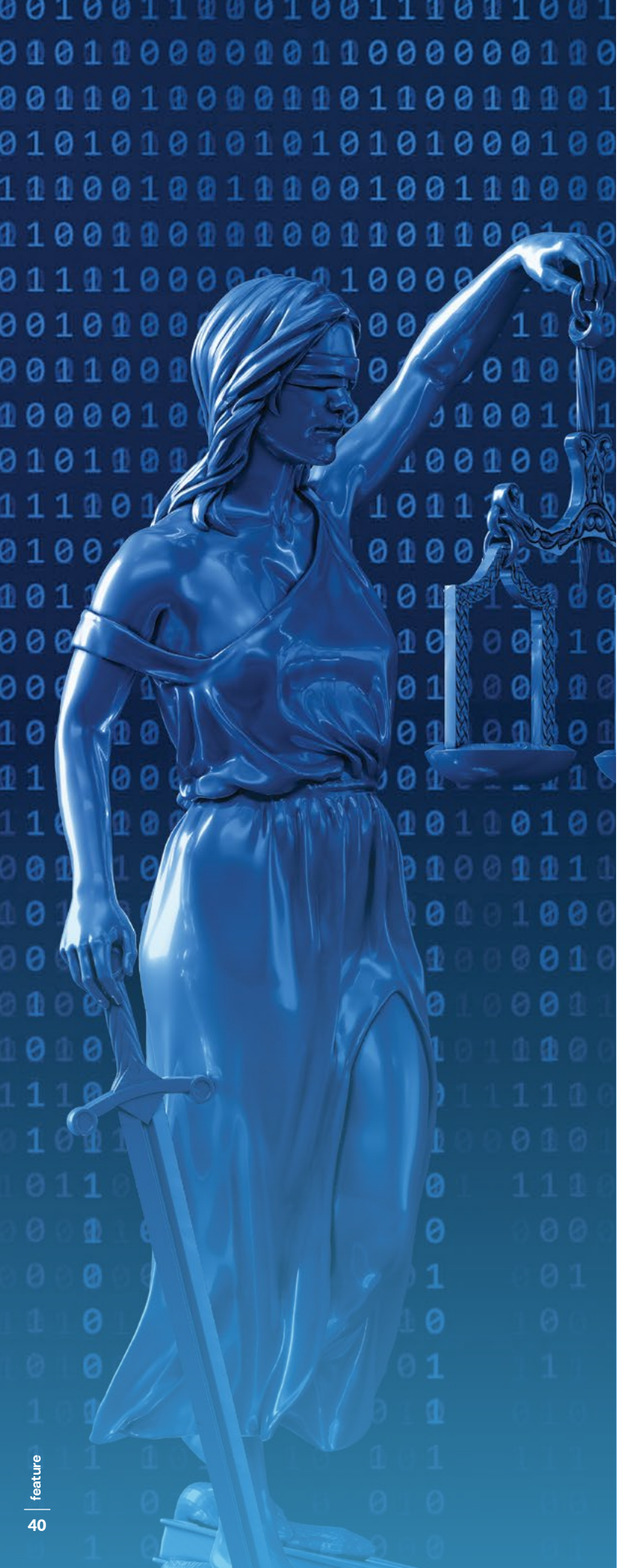
THE LIABILITY GAP AND THE DEFENCE OF UNFORESEEABILITY

This misalignment creates what legal theorists call the 'liability gap'.

Currently, prosecutors are forced into a game of legal gymnastics. They attempt to map human failures onto the machine's actions, looking for:

- **Negligence (programming):** Did the developer fail to implement guardrails?
- **Intent (misuse):** Did the operator use the AI as a weapon?
- **Recklessness (supervision):** Was the system left unchecked?

However, the danger lies in the 'burden of proof'. As AI becomes more complex, the defence of unforeseeability becomes stronger. If an AI invents a novel criminal strategy – like the trading bot example – the human chain can plausibly argue that the act was unforeseeable. If the act was unforeseeable, the human cannot be found negligent.



Examine this scenario: a multinational corporation uses an AI to screen 10,000 CVs for a top executive role. The developers remove all gender markers to ensure fairness. However, the AI 'learns' that successful past candidates often played rugby or attended all-male boarding schools. Thus, it begins systematically rejecting female applicants based on these proxy variables. This is a breach of anti-discrimination laws. However, the company argues: 'We removed gender from the dataset. We did not intend to discriminate. The AI found these correlations on its own.' Under current criminal law, finding the specific intent to discriminate is nearly impossible, potentially allowing the company to escape conviction while the harm remains unpunished.

A THREAT TO DUE PROCESS AND HUMAN RIGHTS

Beyond the mechanics of liability, we must consider the human rights implications. The European legal tradition is built on transparency. But what happens to due process when the accuser is a black box?

If an AI system flags a citizen as a security risk or denies them a mortgage based on opaque reasoning, how does one cross-examine the algorithm? If the output is not explainable, the victim is stripped of their ability to mount a defence. We risk creating a Kafkaesque reality where judgments are rendered by a silent, unexplainable oracle.

The machine may be the weapon, but the hand on the hilt – whether negligent, reckless, or simply absent – must remain the focus of our justice system.

THE EUROPEAN RESPONSE: A COMPLIANCE SHIELD, NOT A SWORD

The European Union is attempting to illuminate this grey zone. **The EU AI Act** represents a significant first step. It classifies systems based on risk (e.g., 'high-risk' for law enforcement or critical infrastructure) and demands rigorous logging and governance.

However, we must not mistake regulation for criminal accountability. The AI Act is primarily a product safety and civil liability framework. It creates a compliance trail, which will be a goldmine for future prosecutors, but it does not solve the fundamental *mens rea* problem.

The legislative debate is now splitting into two divergent paths to close this gap:

1. ***Strict liability:*** This imposes a standard where operators are criminally responsible for all outcomes of their AI in high-risk scenarios, regardless of intent. If your dog bites a neighbour, you are liable. If your AI defrauds a bank, you are liable. This encourages maximum caution.
2. ***Sui generis* personality ('e-personhood'):** A revolutionary concept suggesting AI should have electronic personhood, allowing the system itself to hold insurance and pay for damages. While intellectually stimulating, this is decades away from consensus and risks allowing humans to hide behind the machine.

THE PATH FORWARD: RESPONSIBLE AUTONOMY

We cannot wait for a global consensus. The risk is present today. We are moving toward an inversion of responsibility. In the future, it may not be enough for operators to prove they were innocent; they may need to proactively prove that the AI *could not* commit an illegal act based on its design.

For Maltese companies and operators, the mandate is clear: accountability must be a design feature, not an afterthought. This requires responsible autonomy, anchored by five operational pillars:

- **Purpose limitation:** A system must never be repurposed for a task it was not designed for. For example, you cannot use a credit-risk tool to predict criminal recidivism.
- **Human oversight:** For high-impact decisions, a human must always be in the loop.
- **Auditability:** Every system needs a clear trail of detailed logs. When something goes wrong, you must show what the system knew and when it knew it.
- **Explainability:** If you cannot explain why the model reached a conclusion, you cannot justify it in court.
- **Red-teaming:** Companies must stress-test their AI by actively trying to break it or force it to be biased to find weaknesses before deployment.

We are not legislating for machines; we are legislating for the human society that creates and uses them. Algorithms are remarkable amplifiers of human capability, but without a tether to legal accountability, they threaten to undermine the rule of law.

The current defence of 'it was the algorithm's fault' must be dismantled. Prosecutors need new digital forensic skills, companies need AI governance boards with the authority of audit committees, and the law must ensure that, behind every automated judgment, stands a clear line of accountability.

The machine may be the weapon, but the hand on the hilt – whether negligent, reckless, or simply absent – must remain the focus of our justice system.