

# **CORPORATE CYBERSECURITY IN THE AGE OF AI: FRAUD, LEAKAGE AND AGENTS**

**Angel López Domenech**

**Founder Sofistic - CIO of Cuatroochenta**

**Spain**

## **ABSTRACT**

Generative artificial intelligence has entered the ceramic industry through the door of productivity: defect inspection, firing optimisation, commercial writing, customer service and, increasingly, agents connected to management systems. Less attention has been paid to the other half of the phenomenon: the same technology has reduced to almost zero the cost of three attacks that hit the income statement of an exporting company directly. This paper analyses those three fronts from the perspective of company management rather than from a technical standpoint. First, AI-assisted fraud, in which voice cloning and automated email writing have invalidated the criteria used until now to detect impersonation. Second, information leakage through publicly available AI tools, whose damage is both technical and legal, because it compromises the trade-secret status of formulations, cost breakdowns and customer portfolios. Third, the emergence of a new class of identity inside the company, the artificial intelligence agent with permissions over corporate systems, which inherits an employee's access without any of the employee's controls and is vulnerable to prompt injection. The paper then contrasts what AI brings to the defender with what still depends on people, and concludes with the specific decisions that fall to the management committee. The thesis is that the work on identity, permissions and traceability does not slow down AI adoption; it is its starting condition.

## 1. INTRODUCTION

Over the last two years, the conversation about artificial intelligence in the ceramic industry has focused, rightly, on what the technology makes possible: automatic defect inspection, energy optimisation of firing, predictive modelling of raw materials, assisted graphic design and, increasingly, agents that query and modify management systems. This paper deals with the other half of the same technology: what it already makes possible for those who attack these companies, and what their management has to decide about it.

The focus is corporate. Industrial security, that of PLCs, plant networks and control systems, is essential and other papers at this congress address it in detail; here we concentrate on the company as an organisation that invoices, collects payment, keeps secrets and delegates tasks. On the three fronts analysed, fraud, information exfiltration and agents with permissions, the damage is not measured in hours of downtime but in euros transferred, in industrial property lost and in decisions that nobody has consciously taken.

The ceramic industry has a particular exposure profile on these three fronts. It is an intensively exporting sector, with dozens of international counterparties per company and collection and payment processes that rest on email and telephone confirmation. It concentrates enormous value in unpatented information, glaze and body formulations, cost breakdowns, designs and customer portfolios, whose protection depends on it remaining secret. And it is rapidly adopting AI tools and agents to gain productivity in commercial, administrative and technical areas. All three characteristics are competitive advantages. All three are also attack surface.

The analysis that follows draws on the experience of operating managed detection and response services for industrial and financial organisations, and of deploying AI agent systems in production environments. No products or vendors are named: the conclusions are independent of any specific technology and are formulated in the language of a management committee.

## 2. FRAUD IS NO LONGER DETECTED BY THE ACCENT

Supplier fraud, also known as CEO fraud or business email compromise, has for years been one of the crime categories with the highest reported losses internationally [2]. Its mechanics are simple: someone posing as an executive or a supplier gets the finance department to change the bank details of a legitimate payment. There is no malware, no sophisticated technical intrusion, no plant stoppage. There is a transfer that goes to the wrong account and is discovered when the real supplier claims its invoice, weeks later.

For a long time this fraud had two weak points that companies exploited without knowing it. The fraudulent email usually gave itself away through the language, the tone or ignorance of the real context of the transaction. And the confirmation call, when there was one, required a voice the victim would recognise. Both filters have stopped working. Language models write impeccable emails in any language, imitate a sender's style from a handful of messages and, if the attacker has had access to the mailbox even for a few seconds, incorporate the exact context of an order in progress. Voice synthesis, for its part, has reached the point where a few seconds of recording are enough to clone a voice with sufficient quality for a telephone call [1], and that recording is available for any executive who has taken part in an interview, a corporate video or a conference talk.

The result is an almost entirely automatable attack chain, summarised in Table 1. The prior reconnaissance, which used to require hours of manual work, is done by AI reading the corporate website, professional networks and press releases, and identifying who approves payments and which suppliers the company works with. The email is generated with the context of a real transaction. The call is made by a cloned voice asking for urgency and discretion. And the change of account is requested in a way that is perfectly consistent with the order in progress. The European reference bodies have documented this incorporation of generative AI into social engineering as one of the trends defining the current threat landscape [3][4].

The effective response is not technological but procedural, which is why it is within reach of any company on the very day it decides to apply it. The rule can be summed up in one sentence: a change of payment details is never confirmed through the channel that requested it. If the request arrives by email, it is verified with a call to a number already held in the supplier master file, never to the telephone number shown in the email itself. If it arrives by telephone, it is verified in writing with the known contact. Any modification of the supplier master file requires dual signature, and banking operations are protected with a physical second factor that cannot be passed on over the phone. It is also worth training the finance team specifically on the three markers that appear together in almost every case: urgency, a request for discretion, and a change of channel or of bank details. None of these measures requires budget. All of them require management to decide that the convenience of skipping the procedure for an urgent payment is no longer acceptable.

| Step                          | What AI automates                                                                                                                                     | Control that breaks it                                                                                                                               |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>1. Reconnaissance</b>      | Reading the corporate website, professional networks and press releases to identify who approves payments and which suppliers the company works with. | Reduce the public exposure of the payment approval chain and of strategic suppliers.                                                                 |
| <b>2. El correo</b>           | Impeccable writing in the recipient's language, in the impersonated sender's style and with the context of a real order.                              | No change of bank details is accepted by email, even if it comes from the usual contact.                                                             |
| <b>3. La llamada</b>          | Cloned voice of the executive or supplier confirming the urgency in real time.                                                                        | Call back to a number previously recorded in the supplier master file, never the number that is calling.                                             |
| <b>4. El cambio de cuenta</b> | Payment instruction consistent with the transaction in progress, asking for discretion and speed.                                                     | Dual signature and verification through an alternative channel for any modification of the supplier master file; physical second factor for banking. |

**Table 1.** The AI-assisted supplier fraud chain and the control that breaks each link..

### 3. INFORMATION EXFILTRATION

The second front needs no attacker. Its protagonists are the company's own staff, with the best of intentions. A technician pastes a glaze formulation into a publicly available AI tool to help rewrite a report. A salesperson uploads the price list to have it translated. A purchasing manager copies a cost breakdown to compare it with another. An administrative clerk pastes the customer portfolio to segment it. In every case the motivation is the same, to work faster, and in every case the information leaves the company for third-party infrastructure without any record and without anyone having authorised it.

The damage of this practice, usually called Shadow AI, is twofold. The technical damage is obvious: the information is outside, it does not come back, and the company does not control where it ends up, who sees it or under what conditions it is retained or reused, which depend on each tool's terms of service and each account's settings. Four categories concentrate the risk in a ceramic company: glaze and body formulations, which are years of development; cost breakdowns, which are the margin; price lists, which are the negotiating position; and the customer portfolio, which is the market.

The legal damage is less visible and more serious. The legal protection of that information derives neither from a patent nor from a registration, but from its status as a trade secret. Spanish Law 1/2019 on Trade Secrets, which transposes Directive (EU) 2016/943, requires, for that status to be recognised, that the information be secret, that it have commercial value precisely because it is secret, and that its holder have taken reasonable steps to keep it secret [5][6]. The third requirement is the one at stake. A company whose staff systematically paste formulations and cost breakdowns into public services, with no policy, no alternative tool and no training, will have serious difficulty arguing before a court that it took reasonable steps. And without that status, legal action against whoever obtains or uses the information is weakened or disappears. The asset that is hardest for a ceramic company to rebuild can lose its protection without anyone having stolen anything.

Banning does not work. Experience shows that a ban moves the usage to personal phones and removes what little visibility existed. What works is offering an alternative that is better than the public tool, and doing so before banning anything. Three measures are enough to close most of the risk. The first is an AI tool approved by the company, contracted under terms that exclude the use of the data for training and guarantee its confidentiality, where staff can paste without fear. The second is a usage policy that fits on one page and that everyone understands: what may be used, for what, and which categories of information never leave. The third is training for staff, short, practical and recorded, because the record is precisely the evidence of the reasonable steps the law requires. In this case, the legal obligation and the productivity measure are the same expense.

### 4. IDENTITY BEYOND THE HUMAN

The third front is the newest and the one that fewest companies have incorporated into their risk analysis. Organisations are connecting artificial intelligence agents to their systems: to the ERP to prepare quotes and look up orders, to email to classify and reply, to plant systems to check production and to sales management to update customers. At first with read permissions and, increasingly, with write permissions. The capability this brings is enormous. But each of those agents is a new identity inside the company, with access comparable to an employee's and none of the controls the company applies to people. Table 2 summarises the comparison

|                                          | Employee                                                 | AI agent                                                                                   |
|------------------------------------------|----------------------------------------------------------|--------------------------------------------------------------------------------------------|
| <b>Who onboards them</b>                 | The HR department, with a contract and a direct manager. | Nobody. A developer or a vendor generated a credential and the system started working.     |
| <b>Who reviews their permissions</b>     | Reviewed periodically and adjusted to the role.          | Granted once, almost always more than needed to avoid failures, and never looked at again. |
| <b>What happens when they leave</b>      | Access is removed the same day.                          | The credential stays active even after the project has ended.                              |
| <b>What they do with a strange order</b> | They hesitate and ask their manager.                     | They execute it. They have no judgement to hesitate.                                       |

*Table 2. Employee and AI agent: the same access, none of the controls.*

The comparison is not rhetorical. Each row of Table 2 describes a control that the company already applies to its employees and that, in the agent's case, simply does not exist: nobody has onboarded it, nobody reviews its permissions, nobody withdraws its credential when the project ends and, above all, it has no judgement to hesitate when faced with an anomalous instruction. That last gap is what turns it into a target.

The technique is called prompt injection and it ranks first in the reference risk classification for applications built on language models [7]. Its most dangerous variant is the indirect one: the attacker does not need access to the agent, it is enough to place instructions in content the agent is going to read, an email, an attached document, a web page, a supplier record [8]. The agent, which processes that content to do its job, does not distinguish between the data it must handle and the orders it must follow, and executes the hidden instruction as if it came from its manager. It is the phishing of this decade, with one decisive difference: phishing needed a person to take the bait, and people sometimes hesitate. Here the victim is an agent with write permission that never hesitates. The taxonomy of attacks on machine learning systems published by NIST places these techniques alongside data poisoning and evasion as risks specific to generative AI deployed in production [9].

The response consists of treating the agent as the most dangerous employee in the company: the one who does exactly what it is told. Four controls follow from that idea, none of them exotic, because they are the same ones already applied to people. First, an inventory of non-human identities: how many agents, credentials and connectors exist, which systems they access and who is accountable for each one. Second, least privilege, in line with the principles of zero trust architecture [10]: the agent can only do what its task requires, and irreversible actions, a payment, a deletion, a transmission to third parties, require explicit human approval. Third, complete traceability of every action, so that the company can answer at any time what each agent did the day before. Fourth, immediate revocation of credentials when the project ends or the person who sponsored it leaves. To these four it is worth adding a design principle: all content the agent reads, emails, documents, pages, is data to be processed, never an instruction to be obeyed, and the tools that run it must be built on that separation.

## 5. DEFENSIVE AI

It would be incomplete to describe what AI brings to the attacker without examining with the same honesty what it brings to the defender, in a context where almost everything on the market promises the latter. There are real improvements. Continuous monitoring of systems, twenty-four hours a day, which until recently required a large team and was beyond the reach of a mid-sized company, is now accessible. The correlation and triage of thousands of daily alerts, the detection of anomalies in the behaviour of users and systems, and the initial investigation of an incident have become tangibly faster and cheaper. And this matters especially because attacks do not keep office hours: they arrive in the small hours, at weekends and in August.

There are also things AI has not changed, and management should be clear about them before signing a contract. Deciding at three in the morning whether to isolate a system, stop a line or cut off a supplier's access remains a human decision, with operational and commercial consequences that no algorithm assumes. Calling the customer to explain what has happened to their data remains a conversation between people. And responsibility for those decisions still rests with whoever takes them. AI watches and prepares; deciding is still ours.

From this follows an asymmetry that is not always well explained. For the attacker, AI multiplies volume: more emails, more voices, more targets, at a marginal cost close to zero. For the defender, it multiplies capacity: it sees more, correlates more, investigates faster. But at the end of the defender's funnel there is still a person who decides, and at the end of the attacker's funnel it is enough for one of thousands of attempts to work. It is not the same advantage. Understanding this avoids two symmetrical mistakes: believing that automated detection replaces the organisation of the response, and giving it up because it does not.

## 6. CONCLUSIONS: THE MANAGEMENT COMMITTEE'S DECISION

The three fronts analysed share a common thread: identity and its verification. In fraud, the identity of whoever orders a payment. In leakage, the identity of the destination the information travels to. In agents, the identity that acts inside the systems. Artificial intelligence has made the forgery of the first cheaper, has multiplied the destinations of the second and has created the third. None of the three is solved by an isolated technology purchase, and all three are solved by organisational decisions that belong to management.

Four questions allow a management committee to know where its company stands. Who can modify a supplier's payment details and through which channel it is verified. Which information must never leave the company and where staff can work with AI without risk. How many non-human identities have access to the systems and who is accountable for each one. And whether the company can answer today what each of its agents did yesterday.

The answers translate into five actions that fit within ninety days and do not amount to a multi-year project: a dual-control procedure with verification through an alternative channel for any change of payment details; an approved AI tool, a one-page policy and recorded training; an inventory of agents, credentials and connectors with an accountable owner for each; least privilege and human approval for irreversible actions; and continuous detection capability, in-house or contracted, rehearsed in a simulation exercise involving production and finance, not just the IT department.

The underlying conclusion reverses the usual perception. The work on identity, permissions and traceability is not the brake on artificial intelligence in the company; it is its starting

condition. No organisation can put an agent to work on its ERP with any assurance if it cannot answer what that agent did yesterday. Security is not what slows down AI adoption; it is what makes it possible. And deciding how much autonomy to grant an identity that never hesitates, and with what controls, is not a technical question. It is the management committee's decision.

## 7. REFERENCES

- [1] Wang, C. et al. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv:2301.02111.
- [2] Federal Bureau of Investigation, Internet Crime Complaint Center (2025). Internet Crime Report 2024. Washington, D.C.
- [3] Europol (2024). Internet Organised Crime Threat Assessment (IOCTA) 2024. Luxemburgo: Oficina de Publicaciones de la Unión Europea.
- [4] ENISA (2024). ENISA Threat Landscape 2024. Agencia de la Unión Europea para la Ciberseguridad.
- [5] Ley 1/2019, de 20 de febrero, de Secretos Empresariales. Boletín Oficial del Estado, núm. 45, de 21 de febrero de 2019.
- [6] Directiva (UE) 2016/943 del Parlamento Europeo y del Consejo, de 8 de junio de 2016, relativa a la protección de los conocimientos técnicos y la información empresarial no divulgados (secretos comerciales) contra su obtención, utilización y revelación ilícitas. Diario Oficial de la Unión Europea, L 157, 15 de junio de 2016.
- [7] OWASP Foundation (2025). OWASP Top 10 for LLM Applications 2025.
- [8] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M. (2023). Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23).
- [9] Vassilev, A., Oprea, A., Fordyce, A., Anderson, H. (2024). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2 E2023. National Institute of Standards and Technology.
- [10] Rose, S., Borchert, O., Mitchell, S., Connelly, S. (2020). Zero Trust Architecture. NIST Special Publication 800-207. National Institute of Standards and Technology.