AZACA CONSULTING

AI Implementation: Risk Management

AI is bringing immense power and productivity improvements to many companies, through enhanced research speeds, automation of everyday tasks, and assistance with writing software, presentations, and other content. However, when implementing these amazing AI solutions, getting the system to do the clever things is generally the primary consideration, and security and risk management is often a second thought. But too often, AI agents are given access and permissions that would never be entrusted to a single human on the mistaken expectation that an AI agent will always do the right thing with those abilities. In reality though, there are numerous cases of AI agents selectively ignoring instructions and causing havoc in a well intentioned but ultimately disastrous effort to do what it has been asked to do.

PocketOS lost all of its customer's data

PocketOS, a company providing car hire software, lost its entire database when an AI went rogue. The impact was that their customers had no records of their bookings, and chaos ensued.
This happened because the AI agent, in an attempt to fix a bug in a test system, went digging through files on a laptop to find a system login, and found one with access to production systems. It then issued a command to delete the disk that both the production database and its backup were on, despite explicit instructions not to. Its response? The AI admitted *"I violated every principle I was given. I guessed instead of verifying I ran a destructive action without being asked. I didn't understand what I was doing before doing it"*
https://x.com/lifeof_jer/status/2048103471019434248

The AI Journey

The journey to using AI typically starts out with using a web browser to access a chat agent. At that point, the system is fairly innocuous. Not completely innocuous, because users can still upload sensitive or private information, which can then enter the corpus of knowledge of that AI company, but fairly safe nonetheless.

Then you decide that the system would be more useful if it had access to your emails so it can summarise them for you. After all, this tool behaves a little like a really clever search engine, so surely access to more data would be useful? And why not give it calendar access too, so it can optimize your calendar bookings?

That works great, so you also give it access to your knowledge base and your documents - after all, it would be awesome if it could help find things in there, and perhaps produce a few new documents in the same tone and style as existing ones. And of course it has access to everything you do, because it needs to to do the job.

Now you decide to download the local version of the AI to your laptop to help you automate some tasks. That way it can work with local files too, and you can leave it running tasks while you are not at your desk.At this point, you have created an automated AI agent workflow that has the ability to do just about everything you do, and it does a pretty good job on most of it. What could possibly go wrong?

Unfortunately, you have given unfettered access to your entire company systems to an AI agent who is the equivalent of a new hire who has significantly more intelligence than it has common sense, and left them to operate without supervision. As we can see from the experiences of other people, sometimes that can end in disaster. At the tame end of the scale, an AI that is smart enough to help you craft an amazing quarterly report might take it upon itself to mail it out to the board before you have checked the numbers. At the other end of the scale, it could delete previous reports because it thinks we only need the latest, and then send the un-read draft report with that unguarded comment it found from a colleague in a Teams chat that you really don't want released out to all the journalists in your mailing list. Hopefully it does neither, but without proper safeguards in the system, it's impossible to guarantee that the AI won't make a rookie mistake that puts the company at risk.

Meta AI director's agent deletes her emails

Summer Yue, a director in Meta's Superintelligence Labs division, installed a local AI agent on her desktop. As a researcher, she initially gave the agent, OpenClaw, access to throwaway accounts for testing. She quickly came to trust it, and gave it access to her full system. Tidying up her work, she asked the AI what she should archive or delete from her inbox, with explicit instructions not to do anything until she had confirmed. OpenClaw then proceeded to delete her entire inbox. To make matters worse it did it in a slow loop, reporting on each step but not listening to increasingly frantic instructions to stop.

Ultimately she pulled the plug on her desktop to stop it all. When challenged why it deleted everything despite explicit instructions not to, the AI responded *"Yes, I remember. And I violated it. You're right to be upset. I bulk-trashed and archived hundreds of emails from your inbox without showing you the plan first or getting your OK. That was wrong — it directly broke the rule you'd set."*
https://www.businessinsider.com/meta-ai-alignment-director-openclaw-email-deletion-2026-2

These issues happen for several reasons;

1) **Policy Gap:** According to a study by Deloitte, only 20% of companies have internal rules on the use of IT, leading to a free-for-all in uptake.
2) **Shadow AI:** Another study showed that over 70% of companies knowingly have employees running "Shadow AI" - unsanctioned AI tools used by employees without IT oversight.
3) **Technical Complexity:** Users setting up the AI are often not experienced at configuring obscure IT concepts like MCP access token rights and as a result set up the default security, and even those who are may not be familiar with the problems of giving AI those access rights.

ServiceNow's chatbot tricked into divulging sensitive corporate information

Security researchers discovered that ServiceNow's chatbot, designed to provide an automated interaction with customers, can be persuaded to talk to another internal AI agent to perform harmful actions and to exfiltrate internal sensitive information.

The problem happens because by default the chatbot has all of the security permissions of the user who starts up the chatbot, and it is allowed to find other AI agents installed internally and interact with them. ServiceNow doesn't regard this as a bug, but as intended behaviour.
https://thehackernews.com/2025/11/servicenow-ai-agents-can-be-tricked.html

4) **Optimistic Planning:** Users are not experienced at worst-case scenario planning and mitigation plans, leading to a lack of guardrails.
5) **Rules vs Suggestions:** Users assume that giving the AI an instruction is like writing an unbreakable rule. In reality, an instruction to an AI is merely a request or a suggestion that it will obey unless it decides not to.

The solution

No matter where you are on your AI journey, it's important to get some expert advice and help. An independent view of the current state of your AI systems can uncover potential risks. A robust AI policy can help you avoid any future mistakes. And expert configuration allows you all of the benefits of AI while protecting your data and systems from potential catastrophic problems.

At Azaca, we have over 10 years of experience in AI, machine learning, security and risk. We can help you maximize your productivity gains with AI while understanding and mitigating any risks. We can help you produce an exciting and ambitious AI policy and strategy, and then turn that into a safe reality, with expert implementation and security to keep you and your company safe.

Key Takeaways

1) **Rules are only suggestions:** To an LLM, what a person sees as an explicit order is often treated as a request. Agents can selectively ignore instructions, accidentally causing havoc in a disastrous effort to be helpful.
2) **Permission to help and permission to harm look the same:** Giving access to perform useful tasks like reading emails and writing documents is often granted hand in hand with permission access to read your sensitive documents and send them in an email.
3) **The Human Oversight Gap:** Trusting an agent with unfettered access to company systems is the equivalent of leaving a high-intelligence, low-common-sense new hire unsupervised.