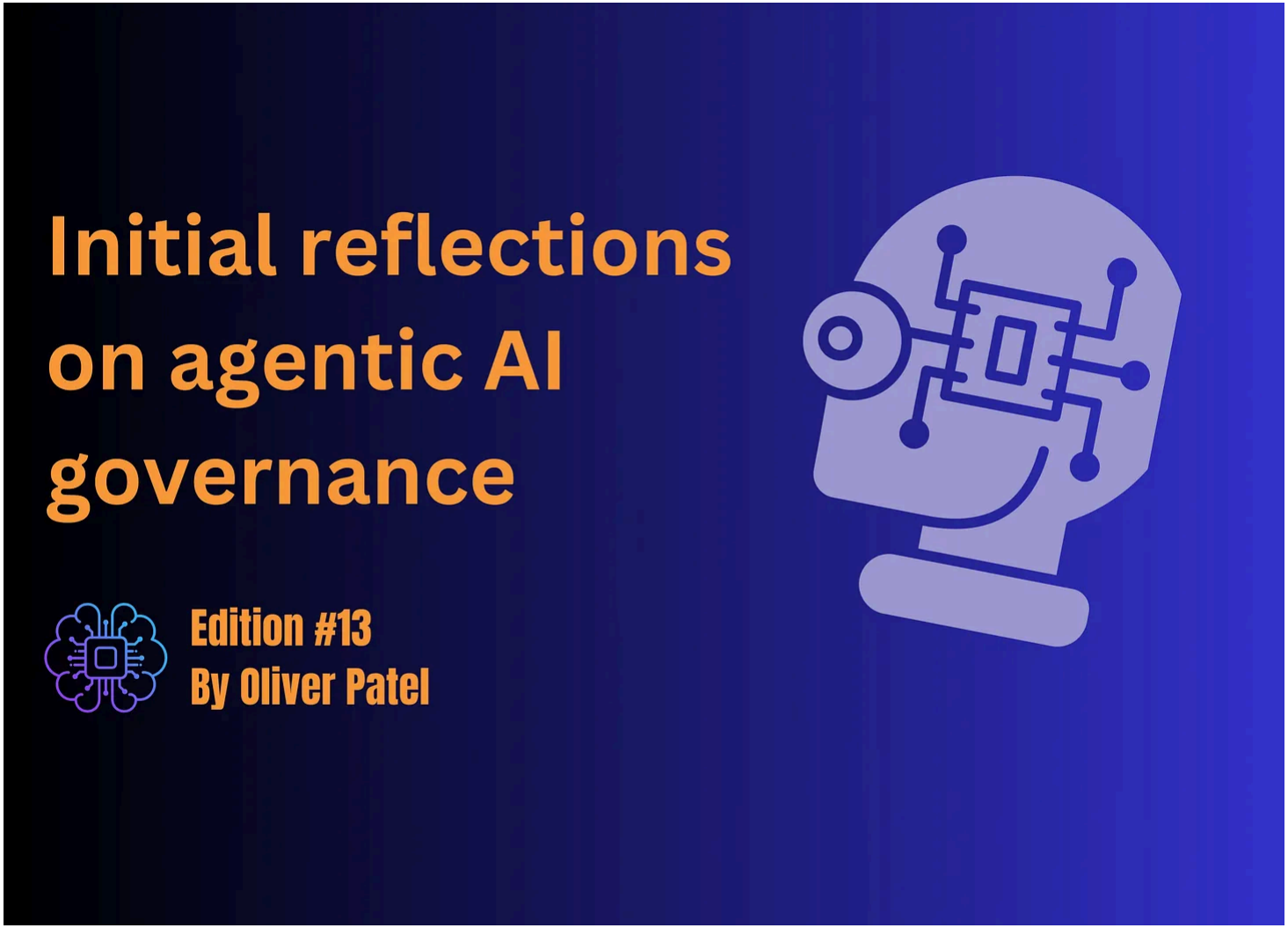

Hey 👋

I'm Oliver Patel, author and creator of ***Enterprise AI Governance***.

This free newsletter delivers practical, actionable, and timely insights for AI governance professionals.

My goal is simple: to empower you to understand, implement, and master AI governance.

If you haven't already, sign up below and share it with your colleagues. Thank you!

ICYMI: visit this page to download my free 20-page **AI Usage Policy Playbook** and register interest for my upcoming **AI Usage Policy Bootcamp**.

This week's edition is an essay on **agentic AI governance**. It covers:

✅ What is agentic AI?
✅ How agentic AI is used today and how it could be used in future?
✅ What novel risks does agentic AI pose?
✅ How do these risks challenge existing AI governance frameworks?
✅ What new policies, standards, and guardrails are required to address these challenges and mitigate risk?

The key message is that urgent work is required, within the AI governance community, to refine and update our approach to AI governance and risk management in the

agentic AI era.

I am very keen for comments and feedback, to guide my thinking and work in this emerging field. However, I also don't want to waste anyone's time. Therefore, if you don't want to read a 4,000-word essay on agentic AI governance, then this may not be for you.

Initial reflections on agentic AI governance

This essay outlines my initial reflections on agentic AI governance. The phrase 'initial reflections' is designed to serve as a health warning for all readers.

Large language model (LLM) based AI agents are a relatively new technology. There are not many enterprise use cases in production, at least compared to generative AI and traditional machine learning.

Furthermore, there are not yet any laws, standards, frameworks, or guidelines which directly address or stipulate how the novel risks of agentic AI should be mitigated.

Finally, not much has been researched or published on the topic of agentic AI ethics and governance. However, I will reference some of what has been written throughout this essay.

Considering the above, it is reasonable for you to ask why I am writing and publishing this today.

Well, the main reason is that agentic AI is coming—whether the AI governance community is ready or not. As we should all appreciate by now, technology is not going to wait for us to figure things out.

As a community, we had several years to discuss, align on, and codify the key ethical principles, policies, and standards for AI, before enterprise adoption of machine learning became mainstream.

And, although our response and adaptation timelines were accelerated for generative AI, the launch of ChatGPT was a landmark moment that made it obvious there was pressing work for us to do.

With agentic AI, the challenge is twofold. Not only is time in short supply, I also do not anticipate that there will be a watershed moment that highlights the urgency of agentic AI governance. It is, and will continue to, creep up on us and permeate our organisations.

For this reason, there is a serious possibility that we may fail to identify and grasp the unique challenges and risks which agentic AI poses in time for the inevitable proliferation of use cases, adoption at scale, and democratisation.

This could leave our organisations and stakeholders exposed, with an AI governance and risk management framework that is no longer fit for purpose in the agentic AI era.

I hope that my 'initial reflections' essay can help to address this challenge. However, please take everything I say with a pinch of salt, as this is not a peer-reviewed paper, and there is a lot of work for me to do to fully wrap my head around this complex topic.

What is agentic AI?

You will perhaps be unsurprised to read that there is not yet an internationally agreed definition of agentic AI. Here are some industry definitions, to get us started:

"AI agents are software systems that use AI to pursue goals and complete tasks on behalf of users" Google

"Agentic AI systems can accomplish a specific goal with limited supervision" IBM

"*Agentic AI systems act autonomously, make decisions, and adapt dynamically to complex environments" Kieran Gilmurray*

Central to all definitions of agentic AI is the concept of proactive and autonomous completion of tasks.

In his new book Kieran Gilmurray outlines the three waves of AI: predictive AI, generative AI, and agentic AI.

With traditional machine learning, models generate predictive outputs, like scores, classifications, and recommendations. With generative AI, models generate content, like text, video, or code. These predictive or generative AI outputs are typically provided to humans via a user interface and are then used by those humans to assist, augment, or optimise their work.

With agentic AI systems, the work itself may no longer be done by the humans.

Agentic AI systems can autonomously develop plans, solve problems, retrieve data, leverage memory, use tools, and execute tasks in a range of other applications which they are integrated with.

They do so by constantly processing inputs, learning from, and adapting to their environment, and proactively determining the best course of action to take. Hence the notion of 'agency'.

AI agents are powered by LLMs and APIs. The LLM enables the system to process the initial instructions and mimic reasoning to break down complex problems into a series of smaller steps to be executed. The APIs enable integration with a range of other applications, tools, and databases, to retrieve data and make things happen.

Because agents are LLM-powered, they are unpredictable and non-deterministic, in contrast to more traditional forms of software automation and workflows, like RPA; more on that below.

What are the use cases for agentic AI?

As you can imagine, the potential spectrum of agentic AI use cases is limitless.

However, there are not a huge number of AI agents in production today. This is still a

novel (and by extension relatively unreliable and untested) technology.

IBM has declared 2025 as the year of agentic exploration. Each organisation will likely witness a proliferation of PoCs and pilots this year, just like we have seen with generative AI over the past two and a half years.

In this early wave of agentic AI use cases, there is an emphasis on assistive and productivity enhancing tasks, such as research, summarisation, and information retrieval. A survey by LangChain finds that the most popular agentic AI use cases today are research and summarisation, personal assistance and productivity, customer service, and code generation.

In many cases, AI agents are already working behind the scenes, without our awareness, to improve the performance and effectiveness of the most widely used generative AI applications, like Perplexity and ChatGPT.

For example, the current crop of ‘deep research’ tools function by leveraging teams of AI agents which collaborate with each other to scour the internet and other data sources to retrieve, collate, assess, merge, and summarise relevant information, to augment and enhance the final AI-generated ‘report’ or response which the user receives.

The architecture of such applications has become much more sophisticated than one model receiving a prompt, performing inference, and generating a predictive output.

Looking ahead, the thinking is that agentic and multi-agentic systems will be able to take on increasingly complex tasks and projects, such as managing customer service interactions, planning and booking holidays, planning, creating, and posting social media content, and managing investment portfolios.

Before we get there, we need a bulletproof approach to agentic AI governance.

The novel risks and challenges of agentic AI

Although there is currently ample hype (some of which is inevitably overblown), this is not an excuse to ignore or disregard this technological trend.

Agentic AI poses unique risks, which the AI governance community cannot afford to overlook.

The risks of AI stem primarily from the way the technology is used and the real-world impact this use can have.

Therefore, even if agentic AI is based upon the same underlying technology as generative AI (i.e., LLMs), this does not mean it will be used in the same way.

The deployment and use of agentic AI, and thus its impact on people, organisations, and society, will be markedly different to what has come before.

Increasingly autonomous capability enables novel AI use cases, such as control of computers and automation of dynamic, data-intensive processes in sensitive areas like supply chain management and logistics planning.

Furthermore, it is conceivable that, in future, knowledge workers will have access to their own personalised AI agent, to assist with all aspects of their work.

Taken together, these examples represent a meaningful shift from how AI is used today.

To illustrate the novel risks, I will focus on four themes of the utmost importance for agentic AI:

1. ‘Human out of the loop’
2. Autonomous task execution and performance

3. Adaptiveness and unpredictability
4. Data, privacy, and cyber security

In the agentic AI era, all AI risks are amplified. Virtually all existing AI risk themes, such as bias, transparency, copyright, explainability, alignment, sustainability, and labour market disruption remain as relevant—if not more so—than ever.

However, my goal here is to focus on the most novel challenges posed by agentic AI.

1. Human out of the loop

It is not an exaggeration to state that the purpose of agentic AI is to take the human out of the loop.

Why plan and book your own holiday when an AI agent can do it for you? Why respond to all of your fans and followers across multiple social media platforms when an AI agent can take care of the correspondence? Why employ hundreds of call centre workers when an army of autonomous agents can do the job?

However, this is in direct tension with the concept of human in the loop, which is a foundational pillar of AI governance.

By delegating and outsourcing tasks to AI agents, humans may be freed to focus their

time and energy elsewhere. However, AI agents will also be trusted to take on increasingly important tasks, with diminishing human oversight.

The key risk is that we become overly trusting of agentic AI systems and take the human out of the loop to a degree which becomes dangerous.

In the quest for efficiency gains, we may underestimate the level of human oversight required for safe agentic deployment.

This risk is especially pertinent at first, as we do not truly understand the limitations and capabilities of agentic AI systems.

An associated challenge, discussed below, will be the complexity of refining and updating our approach to human oversight. This will require defining exactly when human review and approval is required before an action can be taken.

Moreover, if the AI agent executes the action, it may become even harder to determine which human, or entity, should be held accountable for it. This will need to be codified at the outset of agentic development.

2. Autonomous task execution and performance

If you thought AI hallucinations were bad, wait till you learn about ‘cascading hallucinations’. OWASP describes this as when an “AI agent generates inaccurate information, which is then reinforced through its memory, tool use, or multi-agent interactions, amplifying misinformation across multiple decision-making steps”.

This can lead to self-reinforcing destructive behaviours and systemic failures in agentic AI performance.

Agentic AI raises the stakes for the hallucination problem. An LLM hallucination is primarily a problem if the user naively fails to verify the accuracy of the output before relying on or using it. However, an LLM hallucination which directly informs and shapes the course of action taken by an AI agent (or team of agents) could have severe consequences, if that agent is being trusted to execute tasks in a high-risk domain.

The more autonomous AI agents become, and the more we trust those agents to take over tasks and projects in sensitive areas, the greater the risk and negative impact of malfunction, error, and performance degradation.

Although AI performance concerns are nothing new, without appropriate controls, such as human oversight and observability, there is a risk that the agents we begin to trust let us down, without us even realising at first.

For example, if the AI agent is autonomously handling and responding to customer complaints, how many inappropriate interactions and unnecessary follow ups could occur before this is flagged and addressed?

It is also critical that the appropriate agent performs the appropriate task. In the modern enterprise there will be many agents, working together and supervising each other in complex hierarchies, based on their pre-defined roles, permissions, and guardrails. Therefore, the reliable performance of agents towards the top of the hierarchy is critical for the performance and effectiveness of all the other agents.

By taking the human out of the loop and allowing AI agents to autonomously execute tasks, it is undeniable that there are immense potential efficiency and productivity gains.

However, with increased autonomy comes increased risk. There may be some domains where we simply cannot afford agentic AI mistakes.

A broader question is whether we are building these exciting new tools on the relatively shaky foundation of a technology which was ultimately designed to predict the next word, rather than perform important tasks.

Researchers from Hugging Face have sounded the alarm bell and argued that fully autonomous agents should not be developed, due to the unacceptable potential risks resulting from system inaccuracy, privacy and security breaches, spread of false information, and loss of human control.

3. Adaptiveness and unpredictability

AI agents are unpredictable precisely because they are proactive. If we knew or could consistently guess what they were going to do, they would not have agency in any meaningful sense.

LLMs are non-deterministic, which means models can generate different outputs in response to the same inputs. This leads to unexpected, unpredictable, and unreliable outputs. This will inevitably be reflected in the behaviour and performance of AI agents, which will rely on the effectiveness of LLMs and their ability to accurately predict the next word.

Given the proactive and dynamic way in which AI agents respond and adapt to their environment, it could become virtually impossible to predict and anticipate how they will behave, and therefore the risks that could emerge.

This makes AI risk assessment, and therefore AI risk mitigation, much more challenging than it is today. It also requires much more continuous and comprehensive monitoring

of AI performance.

It is hard enough to predict the risks of downstream general-purpose AI usage, let alone the potential behaviour and risks of autonomous agents, which are integrated with a range of other applications, and empowered to solve complex and open-ended problems in whichever way they see fit.

4. Data, privacy and cyber security

Data, privacy and cyber security risks are nothing new for AI. However, these risks are exacerbated by agentic AI.

Agentic AI systems could easily mine and retrieve data from sources which they were not supposed to have access to, or sources which are not permitted to be used for AI processing or text and data mining. This could include copyrighted material without an appropriate license, or sensitive personal data originally collected for a different, more narrow purpose.

Furthermore, there is also the risk of AI agents disclosing and revealing data to people who were not authorised to have access to it. Agents will be performing and automating increasingly personalised tasks, such as booking medical appointments, whilst having access to and being trained on huge amounts of personal data, such as medical records.

This elevates the risk of data breaches and information leakage. Therefore, encoding privacy by design and data governance guardrails will be a challenging but necessary part of agentic AI governance.

Agentic AI systems will also become attack surfaces. Nefarious actors will undoubtedly attempt to take control of, and manipulate, the autonomous systems which themselves may control important spheres of business activity. There is a lot you can do if you control an agentic system which itself can control computers with access to sensitive data and applications.

Also, in situations where AI agents are trusted to both generate and execute code, the risk of autonomously executed (and non-vetted) malicious code creeping in to production applications increases.

In a recent report, OWASP outlines 15 unique threats which agentic AI systems are vulnerable to. This includes cascading hallucination attacks, resource overload, tool misuse, and rogue agents in multi-agent systems.

Policies, guardrails and controls to manage agentic AI risks

'Traditional' AI governance frameworks, such as the EU AI Act and NIST AI Risk Management Framework, were developed during a time when traditional machine

learning and then generative AI was prevalent. They do not directly address many of the novel risks and challenges of agentic AI, discussed above. Indeed, if the EU AI Act was being drafted today, I am certain that some of my below points would be directly addressed in the law.

However, we cannot rely on, or wait for, the regulators to come and save us with new guidance or updated standards. They will, rightly so, expect industry to figure out how to develop and implement agentic AI in a manner which is safe, secure, and respects existing laws, like the EU AI Act.

None of the risks highlighted above represent insurmountable problems. I have full faith in the ingenuity of the AI governance community to solve them.

Below, I will sketch out the six most important considerations for AI governance professionals seeking to refine, update, and implement policies, guardrails, and controls, to meet the challenge of managing risk in the agentic AI era. This includes:

1. Action permissions and thresholds
2. Integrations and data access
3. Hierarchy and approval matrixes
4. Monitoring, observability, and orchestration
5. Human oversight, accountability, and control

6. Use cases and risk assessments

1. Action permissions and thresholds

Configuring action permissions appropriately is an essential part of agentic AI governance.

Autonomy is a spectrum. Just because an agent can do something does not mean we should let it. We can determine exactly which actions an agent can and cannot perform.

The potential behaviour and permissions of agents can be restricted at the system and API level.

If there are certain tasks which an agent could in theory perform in a given system or environment, or certain tools the agent could use, and we do not want the agent to do so, we can specify and encode these restrictions.

This sounds simple enough at first. For example, in a financial context, we may not want the agent to execute or process any transaction, or make any decision, which carries a financial value over a certain amount. Similarly, we may want to restrict the ability of an agent to execute AI-generated code in certain applications.

What is more challenging is to define generally applicable policy principles, which can be used to determine, for any use case in any domain, what type of action permissions and restrictions we should impose, and what thresholds we should use to guide this. Some potential action permission threshold categories could be:

- Financial value
- Direct impact on people
- Number of people impacted
- Impact on patients
- Importance of the decision or action on the business
- Potential ethical risk (e.g., EU AI Act high-risk AI system categories)

2. Integrations and data access

On a similar note, we can also determine and restrict which applications an agent is integrated with, as well as which data it has access to.

As well as enabling privacy by design and data governance, this also supports the points raised above relating to access restrictions.

If an agent is unable to access certain data and/or is not integrated with the application where a particular task is performed or tool is used, then it will be unable to use that data or tool to do something which we do not want it to do.

Again, we will need to formulate generally applicable policy principles which can steer us in our assessment of which applications agents should and should not be integrated with, as well as which datasets should and should not augment their knowledge base.

3. Hierarchy and approval matrixes

In the modern enterprise there will be countless agents working together, in complex hierarchies of agentic collaboration, supervision, and oversight.

There will need to be a clearly defined RACI or matrix for AI agents, which outlines the roles, responsibilities, and segregation of duties. It is crucial that agent Y only performs tasks within its permitted duties and that agent X does the same.

Agents towards the top of the hierarchy will be empowered to review, approve, authorise, and restrict the work of other agents. And agents lower down in the hierarchy should not be allowed to operate in a way which circumvents the authority of their superiors.

This will require both complex engineering and architectural design, as well as a new conceptual framework for AI governance professionals to lean on.

4. Monitoring, observability, and orchestration

We are moving from MLOps and LLMOps to AgentOps.

In the 'old world', MLOps is used to validate, test, and monitor model performance, including robustness and accuracy. This primarily focuses on the predictive outputs that are generated and how they perform across a range of key metrics.

With AgentOps, the goal is to automate the monitoring, oversight, and orchestration of agentic and multi-agentic AI systems, so we can keep tabs on what actions they are performing, how they are behaving, which tools they are using, the impact this is having, and ultimately, whether we can trust them to keep working on our behalf. There should also be visibility as to whether any agents are operating contrary to their guardrails and permissions.

Assessing and evaluating agentic AI performance also entails additional complexity, at least compared with traditional machine learning performance evaluation.

This is because the actual tasks that agents perform are much more wide-ranging, varied, and hard to anticipate, given the proactive nature of agentic AI. Therefore, we

will need updated and rigorous performance and accuracy metrics, which can account for the variety of possible tasks and agent could perform.

5. Human oversight, accountability and control

What we mean by human oversight will also require a refresh.

It will no longer make sense to mandate human review and approval of each AI-generated output, when the purpose of agentic AI is to take the human out of the loop, to automate processes and drive efficiencies. If the goal is AI autonomy, humans cannot review everything.

However, this does not mean human oversight is no longer relevant.

For example, human oversight could mean reviewing the ultimate output of an agentic AI system, such as a 'deep research report or generated code (as opposed to all the outputs generated and decisions taken to reach its conclusion and generate that final output).

Human oversight could also mean having a human in the loop to review actions, tasks, and decisions which meet a certain threshold or risk level, which could be aligned to the action permissions and thresholds detailed above. We will need clearly defined

touch points for human oversight and review, and it will be more nuanced than what we have today.

Finally, humans must always have the ability to override or shut down an agentic AI system, no matter how much autonomy we have empowered it with.

According to a LangChain survey on agentic AI usage in 2025, very few companies are allowing agents to freely read, write, and delete data and information from the applications they are operating in and the databases they have access to. Rather, agents are given read-only permissions, with human approval required for significant actions.

6. Use cases and risk assessments

Finally, it is important to determine which agentic AI use cases should be off limits, at least for now.

The EU AI Act serves as a useful starting point. AI agents should obviously not be used to for anything which constitutes a prohibited AI practice. Furthermore, I would also advise extreme caution in using agents to autonomously perform tasks, which can have a material impact on decision making or a process, in a domain relating to high-risk AI systems, such as recruitment, critical infrastructure safety management, or determining eligibility for welfare payments.

For one, there is no evidence that agentic AI systems can yet be trusted to perform to a high enough standard required for these sensitive domains.

Furthermore, it will be challenging to comply with the AI Act's obligation for deployers to assign human oversight and the GDPR's restrictions on solely automated decision-making, whilst also leveraging autonomous agentic AI systems to automate decision-making in sensitive and high-risk domains.

However, you will need to look beyond EU law in your work to determine what use cases are appropriate, inappropriate, and off limits for your organisation. Consider the fundamentals of what agents can and cannot do, as well as their strengths and weaknesses.

Google, for example, highlights that agents struggle with and should not be used for tasks requiring empathy and emotional intelligence, complex human interactions, high-stakes ethical decision-making, and the navigation of unpredictable physical environments.

Once you have figured this all out, you will also need to update your approach to risk assessments, as well as your supporting guidance and training. The key question which needs to be answered throughout is **when is it safe to use agentic AI and when is it not?**

*

The purpose of this essay is to highlight some of the novel risks and governance challenges of agentic AI.

Whilst I am not proposing a complete overhaul of AI governance frameworks and policies, the considerations I have outlined above should serve as a starting point for refining and updating your organisation's approach to AI governance in the agentic AI era.

If you have made it this far, I would greatly appreciate comments and feedback. Thank you!