AI Autonomous Agents Brainstorming

Scratchpad for determining our initial draft outline

Timeline: end of first quarter est. out of the door by March 31st.

Outline

- Introduction
  - What do you need to know up front?
- Scope / Audience
- etc
- etc
- etc

Points of focus

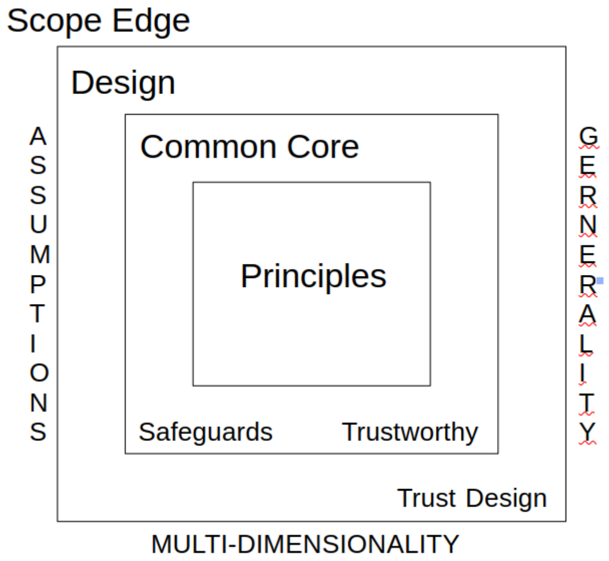
Goal

- Define what an agentic system is
  - Help readers understand agentic systems and what makes them different from previous systems
    - Why
    - How
    - What
  - Present overall high-level architecture and main data flows
  - Concerns that arise that are unique to agentic systems / controls that need to be implemented
    - Risks and their mitigating controls
    - Identify key risks unique to autonomous agents, such as data poisoning, adversarial attacks, and privacy breaches.
- Two main topics we should cover:
  - Data Access
    - Techniques to minimize risk (e.g. data anonymization, see State of Data Privacy Engineering draft for additional ideas)
  - Interactions with other systems/components (aka APIs, DB calls, etc)
    - Inter-agent communication & explainability/observability/traceability (i.e., agent orchestration)
    - Enlist key identity & access aspects related to distributed & automated nature of AI agent systems (eg there is some good & relevant content in CSA LLM Authorisation report that can be referenced and expanded)

- How to design an agentic system to be secure and fulfill business needs (according to the scope we decide on)
- Industry alignment - cross-reference related efforts (eg OWASP LLM top10 relevant vulnerabilities/mitigations)

Mark Y. notes from the 10/30/24 meeting

- Boundaries set by humans; canned routines
- A concern is consumability: the mix of data integration, UI/UX, agent interface
- Best to define scope - where the story starts and stops - what is the crayon boundary? Is it “security of AI agents,” or “the outcome of not securing AI agents.”
- Define what is an agent (e.g. is it ‘system’ or an instance of an agent running, or many instances of agents running comprising a ‘system.?’
  - Define examples, fill in assumptions, use generalization to form boundaries in terms of how to use in business context.
- Wrt to AI agents what is the separation between tasks, tool use, and actions. A human always exist in the loop - humans label, train, execute and monitor running models.
- A visual of boundary: composed from 10/30.2024 meeting



GPT summarization: goals

The paper will define agentic systems, explaining their distinct characteristics compared to traditional systems and detailing their architecture and data flows. Key areas of focus will include unique risks and necessary controls, with specific attention to data access, risk-mitigation techniques (e.g., anonymization), system interactions, inter-agent communication, and explainability. The paper will address identity and access concerns inherent to the distributed, automated nature of AI agents, referencing industry standards where relevant. Additionally, it will establish boundaries for secure design and provide industry-aligned examples to contextualize AI agent use in business settings, emphasizing the continued role of human oversight in managing these systems.

Audience

For this paper (high level, evergreen) - C-Level / everyone
For the related implementation papers - Engineers/developers, hands-on and intended for those building systems directly

Scope

Keep paper as small as possible

- This paper will be kept high level and as small as possible, and reference the specific implementation papers written in parallel
    - Methodology, direction, audience can be C-Level / everyone
    - Specific implementation papers would be directed towards developer community building the systems

- Can write follow-on papers in parallel to capture the more specific guidance
- Keep this paper high level focus, and reference the specific implementation papers instead of going into the weeds
- High level paper would be evergreen, while implementation papers could be updated quarterly to keep up with the fast pace of development

- Two main topics we should cover:
    - Data Access
        - Techniques to minimize risk (e.g. data anonymization, see State of Data Privacy Engineering draft for additional ideas)
    - Interactions with other systems/components (aka APIs, DB calls, etc)
        - Inter-agent communication & explainability/observability/traceability (i.e., agent orchestration)

GPT summarized: scope
Key considerations in agentic system design: It will cover data access, risk-minimization techniques, and system interactions, including inter-agent communication, explainability, and traceability. Detailed technical guidance will be addressed in parallel, regularly updated implementation papers aimed at developers.

Scratchpad

Questions:

- What concerns / anxieties are common when it comes to AI agents?
- How do we define agent autonomy?

    - Kurt Seifried: There are some scenarios where human approval/intervention (human in the loop) is impossible due to the speed of the actions (e.g. high frequency trading)
    - Devon Artis: You wouldn’t give certain permissions to an intern, but what if your autonomous agent has that access? - What controls need to be in place to trust an agent more than a junior admin?
        - Mark Yanalitis: Robotic Process Automation - Autonomous Agent could be treated similar to any other automated process.
        - ?: Lots of different folks splitting tasks along roles, and assigning an AI agent to each. (e.g. crew ai)
            - Adam Lundqvist: Different levels/definitions of autonomy per agent within a team of AI agents
- What level of guidance should we tackle (high/vision level, mid/strategic level, low/code level)
- Nate Lee: How to deal with indirect prompt injection? It’s easy enough to detect “Ignore X and do Y instead” when it comes to overriding the system prompt but much more difficult to detect if it’s the user who changed their mind when the original request and any data that gets mixed with it are all part of the user prompt.
- Nate Lee: What is an agent? Is it any component that does things based on what an LLM tells it to do? Is it the whole system? How do we draw some boundaries around what’s in scope for the paper? Some standard definitions would be very helpful here before diving into securing them
- What existing research is there on this topic?

Initial starting point: “This paper examines the security challenges associated with autonomous agents powered by Large Language Models (LLMs). It highlights potential risks such as data manipulation and adversarial attacks, and discusses basic strategies to safeguard these systems. The goal is to provide a foundational understanding of the security concerns and outline simple measures to enhance the safety of LLM Autonomous Agents in practical applications.”

Jam session slides from merciavanti@gmail.com :
P AIAutoAgents-jamsession01_09042024.pptx

Craig - Could include a definition of the differences between Autonomous, Semi-Autonomous, Not at all autonomous

- MJ - Start at agent level, then dig into details
-

Goal of paper

Enumerate concerns / causes of anxiousness, for each cover high level how to deal with concern, then specific implementation guidance where possible.

- **Why**
- **How**
- **What**

Christopher Byrd - Focus on anxiousness surrounding autonomous agents, how to deal with risks and get benefit of autonomous agents

- Mark Y - Needs to be useful to people who build; something practical for implementers. Paper should help readers build something using AI Autonomous Agents
- MJ - Lots of good material from google we can use as reference
- Keith Pasley - Ethics would be a key concern when it comes to anxiousness around autonomous agents. Data sovereignty, data governance, data privacy etc. API security is foundational to secure AI (secure app to app API calls) and adds new attack surfaces
- Pratixa - Spell out / enumerate the causes of anxiousness surround autonomous agents, e.g. governance, mistakes or attacks, for each cover how to address those concerns. Final section covering specific implementation guidance. three part: What is the concern (and is it real?), high level how to deal with concern, then specific implementation guidance where able.

Nate Lee - Focus on high level patterns for building agentic systems. Then for each:

- Where would you use the various patterns?
- What problems do they solve?
- What are their relative strengths and weaknesses?
- What are the security implications?
- How do you address them?

Aditya Garg -
To ensure the safe and effective deployment of autonomous agents, this paper should focus on balancing architectural design with critical safety principles.
Key considerations must include modularity, redundancy, and hierarchical control to enhance reliability and isolate failures. Safety mechanisms such as fail-safe systems, human-in-the-loop oversight, and real-time monitoring provide robust safeguards against unexpected behaviors. The framework emphasizes traceability and explainability to ensure transparency, while adhering to data privacy, ethical standards, and regulatory requirements like the EU AI Act and NIST frameworks.

We can break the paper into **3 sections** -

- Foundations and Frameworks - Introduce the fundamental concepts, definitions, and architectural principles of autonomous agents.
- Risks, Safety Mechanisms, and Governance - Address the unique risks of autonomous systems and the mechanisms for ensuring their safety and compliance.
- Practical Implementation and Best Practices - Provide actionable guidance for deploying autonomous agents in real-world scenarios.

Christopher Byrd:
Proposed Vision: Promote the responsible use of AI autonomous agents in cloud computing

Proposed Goal: Provide a resource for AI and cloud computing practitioners which provides a foundational understanding and taxonomy of AI autonomous agents, discusses their relationship with cloud computing, outlines associated risks, and includes strategies for enhancing the safety of LLM autonomous agents in practical applications.

[note: implementation guidance and use cases may be a separate paper?]

Ed Sewell - Cover data access as a core theme (#1 risk, access to data and visibility into what the agent is accessing)

(AR) What AA definition do we accept / use in the context of this paper?
Is “nailing machine” a variance of AA: it performs its intended function in the given environment and, hopefully, learns of its own (or other interconnected AAs in the same / similar environments) mistakes?

MJ - Common questions: How to address adversarial attacks, how does human interaction change(?)? What does a “good” agent look like?

Pranav: From a data protection perspective, what data should be kept apart from the agent (e.g. PII, PHI)?

- Keith: Also, how to test for bias
- MJ: Could also be transparency / explainability issue
- Nate: What is “the agent” in this context?

Breaking the paper in three part to address the goal

1. Address introduction of AI Autonomous agents (AAA).
2. Address Risk, governance, challenges and unknowns
3. Address how to approach AAA and define a guidance path to building a vision about your organization.
4. What are the key requirements of building AI autonomous agents?

Audience of paper

- Technical leaders?
    - CIO?
    - CISO (Chief Information Security Officer)
    - Data officers?
    - Data Custodians

    - Data Owners
    - Data Security officers
- Business leaders and decision makers?
- Technical implementers/architects?
    - Need to clearly define implementation - is this code, or high level guidance?

Do we cover healthcare considering how difficult that vertical level? Perhaps we keep it generalized and avoid specifically calling out verticals? - Devon

- Mark Y - Robotics is an example where reusability is difficult, avoid industry mindsets but cover atomic units reusable across multiple industries
- Rajesh K - Can use human safety being an issue as a guide for what subjects to avoid, e.g. healthcare, autonomous vehicles, etc. Just too difficult a problem to solve currently, let those industries address it
- Nate L - I'd avoid vertical specific unless it makes sense to have a brief callout in a given circumstance, there's more than enough general guidance we could give that would be useful that covering vertical specifics would just limit the audience and usefulness.

Scope of paper

Focus on how to take your regular (not autonomous agent) and make it semi or fully autonomous - anxiousness around moving to semi/fully autonomous

MJ Schwenger:
Diff in definitions:

- **Autonomy Level**: Regular < Semi < Autonomous
- **Human Involvement**: Regular (high), Semi (moderate), Autonomous (low)
- **Decision-Making**: Regular (rule-based), Semi (human-assisted), Autonomous (self-directed)
- **Learning**: Regular (none), Semi (limited), Autonomous (continuous)

Pratixa - Section on the types of agents (level of autonomy), anxiousness associated with those levels of autonomy

Nate L: Provide an overview of what an autonomous agentic system is, definitions of the components that make it up (Memory systems, knowledge bases, tools, validators, task decomposition, orchestrators, critics, etc). Discuss types of problems that they are well suited to solve vs. those where they are not and how that will evolve as time goes on.

After overview, focus on high level patterns for building agentic systems. Then for each:

- Where would you use the various patterns?
- What problems do they solve?
- What are their relative strengths and weaknesses?
- What are the security implications?
- How do you address them?

Rajesh - Significant real fear/pushback from customers when it comes to autonomous agents. Knee jerk reaction that it's scary/bad. Articulate and leave to customer to decide?

Keith - In the middle, not high level, not specific technical implementation (aka code), in-between to bridge the gap. Intended for systems architecture, how do you design the systems around an AI autonomous agent

- Pranav:
    - Build one document for high level CIO - leadership buy-in
    - Second document for implementers - technical guidance
    - Akshay title ideas:
        - Paper 1: Strategic Implications of AI Autonomous Agents
        - Paper 2: Building Secure AI Autonomous Systems

Keith - Awareness document or best practice document?

Devon Artis - Should this be called a Security Implementation Guide? Or is this a high level focus?

- Devon - e.g. if we want to hand this paper to an engineer to let them build, what do we need to write?
- Josh Buker - Would definitely like something practical we could hand to devs/engineers

Adam: Lots of existing papers, make sure that what we're writing is covering new ground. Focus on cloud aspects?

- Pratixa - Would be good to include those other papers as references, and also use them to do a gap analysis when it comes to research around AI autonomous agents. Give a current state of as a section of the paper?
- Pranav - There might be bias introduced looking at other research
- MJ - lots of high level guidance
    - AL: (e.g. from McKinsey, etc)

Papers by CSA that could be valuable resources for us:
https://cloudsecurityalliance.org/artifacts/securing-llm-backed-systems-essential-authorization-practices (quite technical, rather for architects)
https://cloudsecurityalliance.org/artifacts/using-ai-for-offensive-security (more high level, but many references to concrete implementations)

Is this technical details or techniques and algorithms / cybersecurity implications? (MJ Schwenger)

~~Healthcare related AI agents can be a specific focus - especially in regards to liability~~ (Rajesh Kanungo) - So many issues we might NOT want to address WRT healthcare, hard to apply our guidance considering the complexity of healthcare regulation/liability

- Perhaps more generic/multiple verticals? (Govindaraj)

- Similar thoughts from Devon? (I missed convo while notetaking)
- Are the verticals our specific focus, or what is our core goal with the paper from the CSA perspective? - Adam
- Autonomy and control mechanisms with Human oversight or Human-in-the-Loop (HITL) and Fail-Safe mechanisms

Usage of AI Agents within the cloud, or do we cover edge agents as well? (Ed Sewell)

- Electrical grids and renewable energy is one area looking at autonomous agents for rollout
- Ed Sewell interested in being a Lead Author as well

James Swift - Should embedded Systems be included?

Devon Artis - Estimated Delivery Date of a draft

Ken Walling - Section covering current and proposed legislation that could impact autonomous agents. e.g. kill switches in *models* ~~autonomous vehicles(or agents?)~~ New CA law doesn't seem to be specific to vehicles - but they would be a likely impact: link to ars Technica article

…"As we've previously explored in depth, SB-1047 asks AI model creators to implement a "kill switch" that can be activated **if that model** starts introducing "novel threats to public safety and security," especially if it's acting "with limited human oversight, intervention, or supervision."

- Mark Yanalitis - Probably will have a human in the decision loop for several years at minimum.
    - Still have yet to use AI for self referential data/decision validation

Do we cover third party integration risks? (Govindaraj Palanisamy)

Security Concerns:

- **Vulnerabilities and Threats:** Identify potential security risks associated with autonomous agents (e.g., adversarial attacks, data privacy breaches).
- **Defense Mechanisms:** Discuss techniques for protecting autonomous agents from attacks (e.g., anomaly detection, intrusion detection).
- **Ethical Considerations:** Explore ethical implications related to autonomous agent behavior, anomaly detection and decision-making. (I guess this should not be in security concerns section?)
    - Bias and Fairness: Address the ethical implications of bias in autonomous agent decision-making and how to mitigate it.
    - Accountability and Transparency: Discuss the need for accountability and transparency in autonomous agent systems, particularly in high-stakes applications.
    - Privacy and Data Protection: Section to examine the privacy and data protection concerns associated with autonomous agents and how to address them.

- **Attack Surfaces:** Analyze the attack surfaces of autonomous agents, including their software, hardware, and communication channels.
- **Fallback Mechanism** - ensure that the agent has a clear fallback mechanism or procedure, in case it encounters a scenario it can handle.
- **Regulatory frameworks:** Explore existing and emerging regulations governing the development and deployment of autonomous agents.
- **Third-party dependencies** - if your agent depends on 3rd party apis - ensure that the 3rd party tool is vetted
- **API security** -Keith Pasley-discuss fundamental requirements
- Access to PII, PHI (Personal Identifiable information) for automated processing
- Fraud and Risk Management - in situations of automated access to API for lookup of information handling Fraud and Risk is very important.

(AR) Conventionally, threat modeling is a very resource intensive and futile undertaking, unless the set of objectives is limited to those critical. Since TM is an iterative process, the set can grow, as priorities shift.
When you talk about vulnerabilities, you imply the environment. Hence the latter must be well understood and adequately described. Environment should define pertinent threats.

Suggesting taking out ethics discussion, since the entire purpose of AA might be unethical.

Attack surface is environment and system specific: it makes little sense to spend resources on petty human physical attacks on SCADA, unless the subject is an insider.

As to the fallback, provided sufficient resiliency is built in to uphold the intended functionality, it is unclear. Ex.: a mine clearing AA can withstand only a certain mine capacity, say, 1.5 kg. And there is no warning attached: if you are a max 1.4 kg AA, please, stay away. There is no sensing / perception capability that might support desired resiliency.

-

Vaibhav Malik-

Should add some things like:

Risk Assessment Framework: Develop a specific risk assessment framework for AI Autonomous Agents, incorporating elements from standard risk management methodologies but tailored to the unique challenges of autonomous systems.
Use Case Analysis: Include detailed use case analyses that bridge the gap between theory and practice. This could cover various sectors like healthcare, transportation, and critical infrastructure, focusing on how autonomous agents can be implemented securely and effectively.

API Security Section: As suggested by Keith Pasley, include a comprehensive section on API security, discussing fundamental requirements and best practices for securing the interfaces that autonomous agents use to interact with other systems.
Regulatory Compliance Guide: This guide provides an overview of current and proposed legislation relevant to autonomous agents, including guidelines on how to ensure compliance in different jurisdictions.
Architectural Patterns: Discuss various architectural patterns for implementing autonomous agents securely, including considerations for centralized vs. decentralized approaches, and microservices-based architectures.
Data Access and Visibility Controls: As Ed Sewell emphasized, include a detailed section on data access controls and maintaining visibility into what data the agent is accessing, as this is considered a core risk.
Threat Modeling for Autonomous Agents: Develop a specific threat modeling approach for autonomous agents, focusing on their unique attack surfaces and potential vulnerabilities.
Fail-Safe Mechanisms and Human Oversight: Elaborate on the implementation of fail-safe mechanisms and human-in-the-loop (HITL) processes to ensure safe operation of autonomous agents.
Performance Metrics and Evaluation: Include a section on how to measure and evaluate the performance and security of autonomous agents, providing concrete metrics and benchmarks.

(MJ) If we will work on tech details:
Technical Foundations

- Agent Architecture: Discuss various agent architectures (e.g., reactive, deliberative, hybrid) and their suitability for different applications. Centralized vs. Decentralized: Discuss the advantages and disadvantages of different architectural approaches.
- Machine Learning Algorithms: Explain the underlying machine learning algorithms (e.g., reinforcement learning, deep learning) that power autonomous agents.
- Decision-Making Processes: Describe how autonomous agents make decisions, including planning, reasoning, and learning. The coverage of decision transparency and auditability
- Perception and sensing: Explore the technologies used for agents to perceive and interact with their environment (e.g., sensors, computer vision).
- Communication Protocols: Discuss how autonomous agents communicate and interact with each other and their environment.
- Action Execution: Describe how agents execute actions in the real or virtual world (e.g., actuators, control systems)

Ed Sewell - Risk framework?

MK: Proactive Operations scope for Technology to use the signals and take actions to deliver proactive stability to platforms/Systems that are critical to society.

Example Use Cases:

- **Healthcare:** Predict equipment failures in medical devices to prevent disruptions in patient care.
- **Transportation:** Detect anomalies in traffic patterns to optimize traffic flow and prevent congestion.
- **Critical Infrastructure:** Monitor power grids for signs of instability and automatically initiate preventive measures to avoid outages.

Key Requirements:

- Clean data source
- Human oversight without introducing toil or noise.
- Ethical and Regulatory compliance.

Are we focusing on specific use-cases? - Manas Khanna

- Mark Y - Use cases can be useful because it helps bridge the gap between theory and practice in a natural way
- MJ - Can also provide a feedback loop on high level theory and how to implement
- Mark Y. - TL;DR, yes. We will focus on use-cases.

Gokhan Polat - Can we categorize types of autonomous agents?

- Mark Y -
- Chris B - Agreed on categorization by function
- Ken Walling - riffing on Gokhan's notion - perhaps we should also define the difference between General AI and Autonomous Agents.
- Devon Artis - Agree with a clear definition

Keith Pasley - API Security

Chris - My 2c - there is room for a general document with the taxonomy, risk, safety discussion, and then another paper for implementation guide

Devon Artis - Should we be calling this a security implementation guide ? That would help with what should be written

Ed Sewell - Keep focus on (data) security, protocols, reference architectures that can be implemented, etc. Use that as a starting point. As far as risk mgmt frameworks, e.g. NIST, EU AI Act, etc (reference them in our paper?)

MJ - Bank of America has a well accepted implementation, interesting to look at perhaps?

Adam L - Microservices, large agents can fail at tasks, small agents in concert similar to microservices could avoid those issues in the short term until large agents

Comment from Alex Getsin:

- To keep with our working group scope (AI Tech & Risk), I would stray away from ethical considerations or governance unless they context appropriate to something we discussing from a tech/risk perspective.
- I would also avoid discussing 'security landscapes surrounding' and hyper-focus on threats, tech, etc

Alex Rebo:

- Risk management looks very different for different orgs or even individuals - Perhaps draw a line at threat model, because we can easily agree on threats, but not necessarily risks
- Alex Getsin: we can base our scope on owasp and mitre threats and risks, and we can use the extensive https://airisk.mit.edu/
- Alex Rebo: A paper on AA taxonomy: https://lia.disi.unibo.it/corsi/2007-2008/SMA-LS/papers/4/agentorprogram.pdf

Kurt Seifried

- Potential bias in how we're selecting scope: We inherently trust agents that tell a human what to do more than an agent that affects the real-world directly.
- Candy Alexander: This is likely about the fear of risk impact with "real world" - aka safety of life

Ken Walling: We shouldn't avoid things, instead outline the scope at the top and our reasoning at the beginning of the paper.

Akram Sheriff:

- In agentic workflows, Large Language Model (LLM) agents are designed to autonomously accomplish tasks or achieve specific goals by interpreting and acting upon user inputs. However, a critical issue arises in the form of a "trust problem" at runtime.
- This trust problem manifests when the LLM agents are either unable to validate or insufficiently trust their own outputs, external data sources, or the sequence of actions taken during task execution. This should be also covered in a detailed manner.

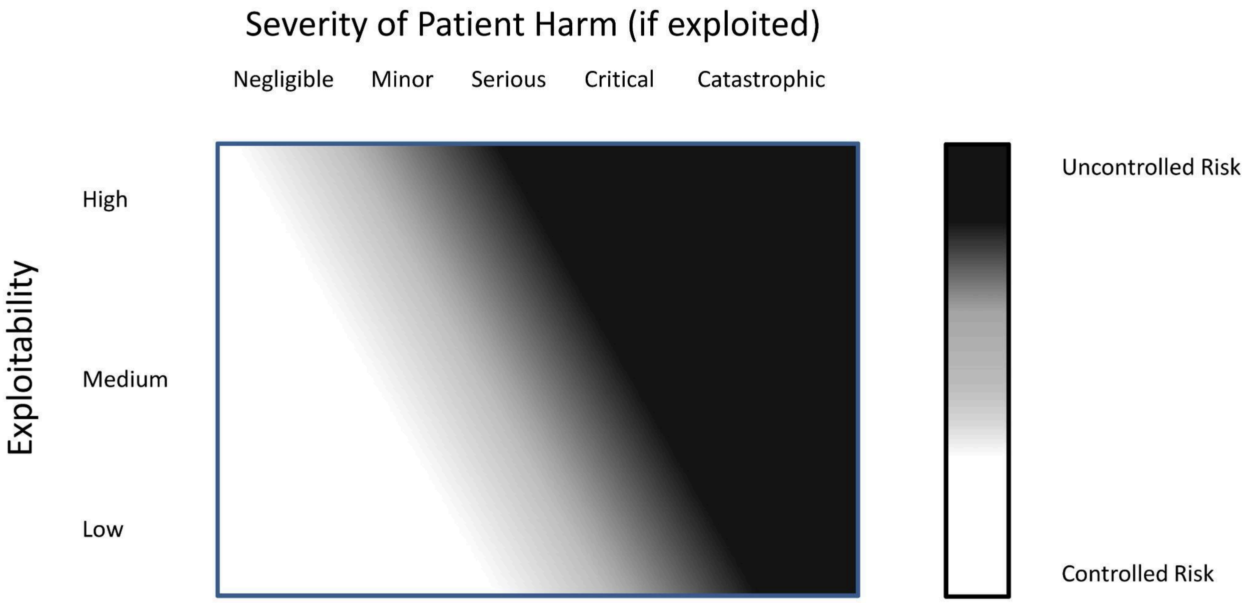
Candy Alexander:

- Definition (with examples)
- Assumptions of direction
- Scope
- Broad stroked and more focused to types of agents
- Purpose
- How to secure
- considerations for use/development

Safety and AI Boundaries Reference

https://www.fda.gov/media/95862/download

Setting boundaries on what CSA will address. The FDA has given guidelines that can transcend the medical industry and can be applied to other domains like robotics



Building Trust into a GenAI System in Healthcare: lessons that may be transferrable

Add information about the process used for a non-profit group building the Onco Assistant, a GenAI solution for providing cancer therapy to patients diagnosed with breast cancer

1. Expert in the loop: an experience breast cancer surgeon was introduced early in the process and he brought in
    a. Credentials
    b. Experience
    c. Ability to identify missing pieces or errors
    d. Ability to identify real problems that need to be solved in the field.
2. Stepwise scientific experiments that the FDA promotes
    a. Stage 0: Early experiments: Use publicly available, de-identified data of patients
    b. Stage 1: Work with a doctor. Find 2 study subjects: allows for basic model validation, useability and identification of missing areas including strengthening explanations, references, local policies, procedures
    c. Stage 2: Run the system against 500 questions from oncology board exams: score 94-96%
    d. Stage 3: Increase number of study subjects to 10
    e. Apply for a formal study approval by a medical research hospital.

3. Formal Study:
    a. Will collect data on 40-50 study subjects
    b. Find errors and devotions from the norm
    c. Identify areas where the system is weak
    d. Address weaknesses: Remediate, create guard rails, or create advisories
4. Clinical Study: TBD

All through the study process, all the data, the errors, the remediations, the doctor notes,, and the successive refinements are available.

It is not just the ability to explain the reasons, it is the ability for us to show how we made the system learn to take care of issues, with a doctor's oversight, which will carry the day.

EU Restrictions on AI: Different Rules for different risks

The EU AI Act is useful for defining risk
https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

This is a verbatim copy from their site.

AI Act: different rules for different risk levels

The new rules establish obligations for providers and users depending on the level of risk from artificial intelligence. While many AI systems pose minimal risk, they need to be assessed.

Unacceptable risk

Unacceptable risk AI systems are systems considered a threat to people and will be banned. They include:

- Cognitive behavioural manipulation of people or specific vulnerable groups: for example voice-activated toys that encourage dangerous behaviour in children
- Social scoring: classifying people based on behaviour, socio-economic status or personal characteristics
- Biometric identification and categorisation of people
- Real-time and remote biometric identification systems, such as facial recognition

Some exceptions may be allowed for law enforcement purposes. "Real-time" remote biometric identification systems will be allowed in a limited number of serious cases, while "post" remote biometric identification systems, where identification occurs after a significant delay, will be allowed to prosecute serious crimes and only after court approval.

High risk

AI systems that negatively affect safety or fundamental rights will be considered high risk and will be divided into two categories:

1) AI systems that are used in products falling under the EU's product safety legislation. This includes toys, aviation, cars, medical devices and lifts.

2) AI systems falling into specific areas that will have to be registered in an EU database:

- Management and operation of critical infrastructure
- Education and vocational training
- Employment, worker management and access to self-employment
- Access to and enjoyment of essential private services and public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Assistance in legal interpretation and application of the law.

All high-risk AI systems will be assessed before being put on the market and also throughout their lifecycle. People will have the right to file complaints about AI systems to designated national authorities.

Transparency requirements

Generative AI, like ChatGPT, will not be classified as high-risk, but will have to comply with transparency requirements and EU copyright law:

- Disclosing that the content was generated by AI
- Designing the model to prevent it from generating illegal content
- Publishing summaries of copyrighted data used for training

High-impact general-purpose AI models that might pose systemic risk, such as the more advanced AI model GPT-4, would have to undergo thorough evaluations and any serious incidents would have to be reported to the European Commission.

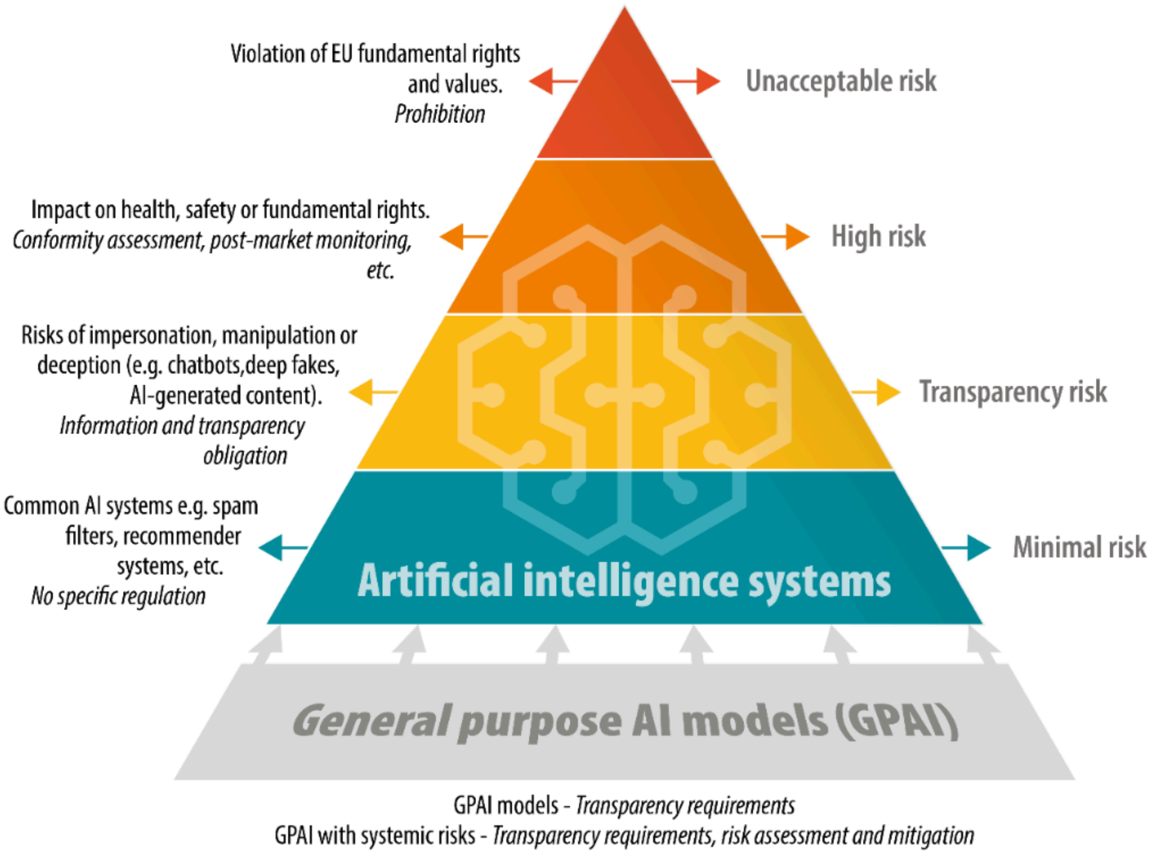
Content that is either generated or modified with the help of AI - images, audio or video files (for example deepfakes) - need to be clearly labelled as AI generated so that users are aware when they come across such content.

AI Agent Safety

References

1. European EU AI Act -https://artificialintelligenceact.eu/
2. NIST 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

Types of Impact


Violation of EU fundamental rights and values.
Prohibition
Unacceptable risk
Impact on health, safety or fundamental rights.
Conformity assessment, post-market monitoring, etc.
High risk
Risks of impersonation, manipulation or deception (e.g. chatbots,deep fakes, AI-generated content).
Information and transparency obligation
Transparency risk
Common AI systems e.g. spam filters, recommender systems, etc.
No specific regulation
Minimal risk
Artificial intelligence systems
General purpose AI models (GPAI)
GPAI models - *Transparency requirements*
GPAI with systemic risks - *Transparency requirements, risk assessment and mitigation*

Does it Impact physical systems?

1. Medical devices
2. Electrical grid

3. Chemical, biological, radiological, or nuclear (CBRN) weapons products
4. Environmental Impact

Can it cause physical harm to human beings or assets?

Examples

1. Turn off insulin injection?
2. Can it cause a dam to spill?
3. Can it shut off the power?
4. Cause a car crash?

What kind of Harm?

We will use the EU AI Act as a source

1. Prohibited
2. High
3. Medium
4. Low

Are there regulations, laws, and standards?

Examples are:

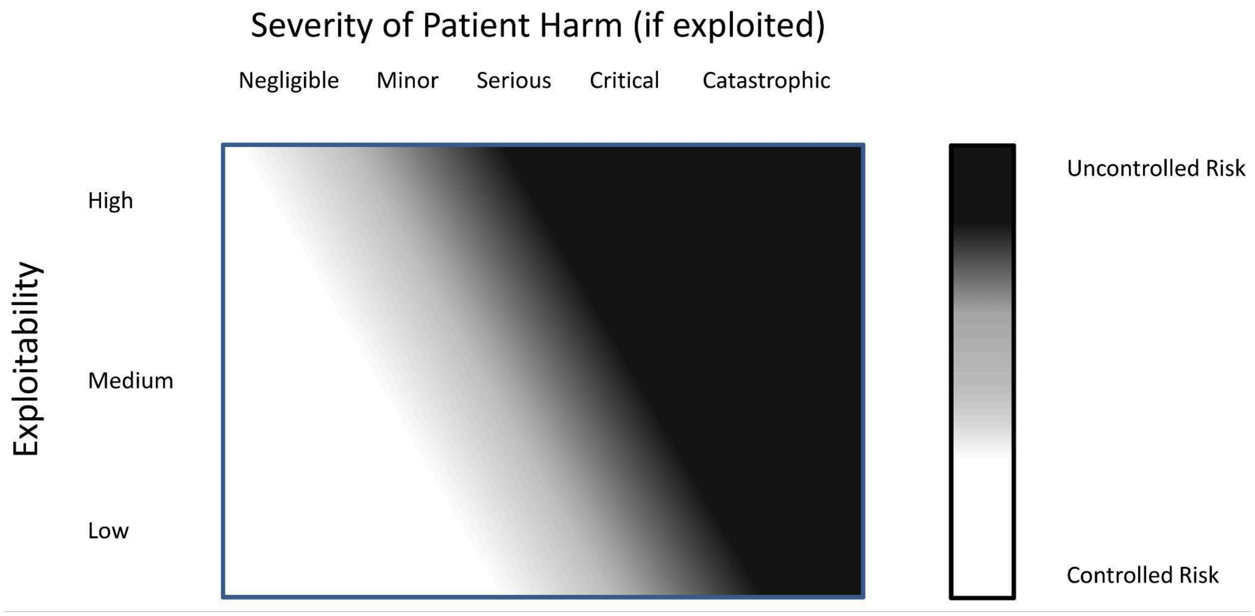
1. FDA for the US
2. EU-MDR for Europe
3. Chemical industry Regulations
4. FAA rules and regulations

Decision Tree

1. If there are regulatory standards, defer to them.
    a. Corollary: if the field is subject to regulations and governance, and GenAI regulations haven't been decided upon, STOP
2. If there are minimal regulations in the domain, use classification schemes to determine potential harm using the following or other similar guidelines.
    a. EU AI Act
    b. NIST 600-1
3. Does the system have adequate protection to prevent harm?

4. Does the service or device provider comply with the
5. Use the FDA heat map and the EU AI Act as guides if regulations do not exist.

[Use Mermaid]
You can have safety with a safe system. You can’t have a safe system with an unsecure system. [Kurt]



What do we do in Response

High Risk

1. Conformity Assessment
2. Quality and Risk Management
3. Registered
4. Post Market Surveillance

Medium Risk

Transparency requirement.

1. Disclosing that the content was generated by AI
2. Designing the model to prevent it from generating illegal content
3. Publishing summaries of copyrighted data used for training

Minimal Risk

How / what we can proceed researching and writing?

A.

if we double-down on the **'RISK'** part, then our writing can discuss a lot of the common risk management components from standard methodologies;

1. Risk ID – A unique identifier for each risk.
2. Risk Description – Detailed description of the risk.
3. Risk Category – Classification of the risk (e.g., financial, operational, compliance).
4. Likelihood (Probability) – Estimated probability of the risk occurring.
5. Impact (Severity) – The potential impact if the risk occurs.
6. Risk Rating (Score) – Calculated risk score based on likelihood and impact.
7. Risk Owner – Person responsible for managing the risk.
8. Mitigation Strategy – Actions or plans to reduce or eliminate the risk.
9. Residual Risk – The remaining risk after mitigation strategies are applied.
10. Response Plan – How the risk will be handled (e.g., accept, avoid, transfer).
13. Control Measures –…
14. Contingency Plan – ,,,

B.

we can also choose one, such as 3. Risk Category or even more specific - top 10 risks for ai autonomous agents.

Devon sample outline

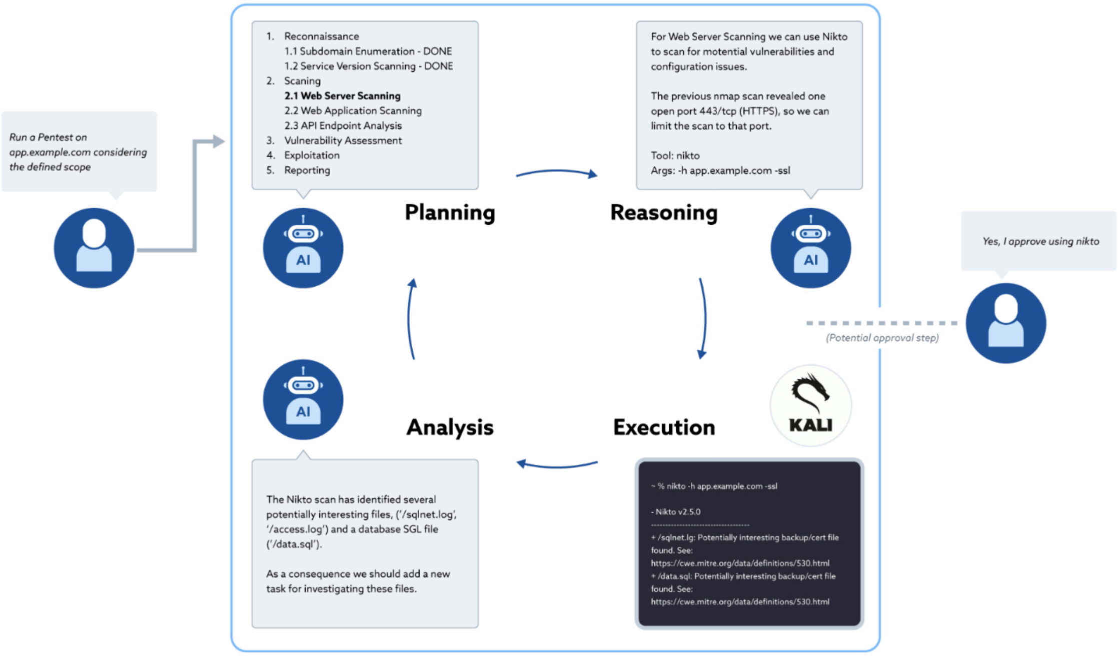
1. Intro to AI Autonomous Agent
2. Types of AI Autonomous agents
   a. Task-Specific
   b. Learning Agents
   c. Mult-Agent System
   d. Real-Time Agents
   e. Mapping Agents to Risk categories using ATLAS
3. AI Risk Framework and ATLAS Threat Modeling
4. Security Considerations of Auto Agent in office and computer system
5. Security Considerations of Auto Agent in Medical System

Adam Lundqvist - possible agent definition (from OffSec Paper):

AI Agents

AI agents are autonomous or sometimes semi-autonomous systems designed to perceive their environments and act to achieve set goals, thus shaping future interactions with the environment. These agents can use the power of LLMs to plan tasks, trigger task execution, make decisions, and interact meaningfully with the world. Unlike basic LLM applications, an AI agent using LLMs follows a cyclic approach to achieve its end goal, continuously learning and adapting from its findings and adjusting its approach. This iterative self-adaptation makes the agent effective at solving complex problems through a multistep process until the task is completed.

An agent begins by breaking down the user request into actionable and prioritized plans (**Planning**). It then reasons with available information to choose appropriate tools or next steps (**Reasoning**). The LLM cannot execute tools, but attached systems execute the tool correspondingly (**Execution**) and collect the tool outputs. Then, the LLM interprets the tool output (**Analysis**) to decide on the next steps used to update the plan. This iterative process enables the agent to continue working cyclically until the user's request is resolved.



Run a Pentest on app.example.com considering the defined scope

1. Reconnaissance
   1.1 Subdomain Enumeration - DONE
   1.2 Service Version Scanning - DONE
2. Scaning
   2.1 Web Server Scanning
   2.2 Web Application Scanning
   2.3 API Endpoint Analysis
3. Vulnerability Assessment
4. Exploitation
5. Reporting

For Web Server Scanning we can use Nikto to scan for motential vulnerabilities and configuration issues.

The previous nmap scan revealed one open port 443/tcp (HTTPS), so we can limit the scan to that port.

Tool: nikto
Args: -h app.example.com -ssl

Yes, I approve using nikto

```
~ % nikto -h app.example.com -ssl

- Nikto v2.5.0
---------------------------------
+ /sqlnet.lg: Potentially interesting backup/cert file
found. See:
https://cwe.mitre.org/data/definitions/530.html
+ /data.sql: Potentially interesting backup/cert file
found. See:
https://cwe.mitre.org/data/definitions/530.html
```

The Nikto scan has identified several potentially interesting files, ('/sqlnet.log', '/access.log') and a database SGL file ('/data.sql').

As a consequence we should add a new task for investigating these files.

Figure 2: AI Agent Phases

Notes: Autonomous Agent Fears and Pushback

This section is a scratchpad based on my company's observations of customer reactions

Agents acting on any critical operational system

This fear has both an irrational component and a rational component.
Customers already use automated processes. The whole area of control theory, especially non-linear control theory, is based on building complex mathematical models. The good thing is

that control systems, especially non-linear ones, (generally) have strict guardrails. The 737-MAX is a counter-example.

We found that customers who could design the guardrails themselves or introduce a human into the system felt more comfortable.

In general, all well-designed critical systems have independent safety monitoring systems.

NASA: Apollo missions
A good source of materials is NASA. I'll look up the documents but they had practices in place to protect against failures at any level. They would have multiple levels of monitoring and safety systems. The monitoring and safety teams were designed by a different team(s).

Conclusion: for agents to succeed in critical infrastructure roles, agents must be trustworthy.
For them to be trustworthy, there must be provable guardrails at the very least that are not built by the same people/technology.

Fear of being replaced

We found that the number of CSOs pushing back against cybersecurity agents was close to 100%. Since we focus on only cybersecurity, we don't have data in other areas.

Humans have designed themselves into processes and procedures. I'll let psychologists handle this one.

Red Team/External Verification Proof

The agent implementation team claims that the agent works perfectly. Customers will not believe it till they see independent verification.

Theory vs. Proof (Field Testability)

Customers were scared of agents may cause harm to their system under unforeseen circumstances.
You have to be able to demonstrate to the customer that no harm could occur: this is impossible to do with complete infallibility.

Other fears: Agents getting too much access, agents being spies (yup),

LLM Agents for doing CSPM and LLM Agents Architecture References: (For Cloud and SAST Security usecases) - Akram Sheriff

https://outshift.cisco.com/blog/intent-driven-llm-agents-enterprise-IT

https://github.com/akramIOT/GENAI_AWS_CSPM_Security_Threat_Agent/tree/main

https://github.com/akramIOT/SAST_ML_TECHNIQUES

https://github.com/akramIOT/CSRF_Security_LLM_Agent

https://github.com/akramIOT/IOT_SECURITY_THREAT_DL

https://docs.google.com/document/d/19seBB_9i8ifG6yPvUzmeQeSkDKuOQGTVLW-0oPWG25M/edit?tab=t.0#heading=h.koewyciymtmx-Agentic AI red teaming guide by Ken Huang

https://cloudsecurityalliance.org/blog/2024/12/09/from-ai-agents-to-multiagent-systems-a-capability-framework

\
https://github.com/kenhuangus/Artificial-Intelligence-Vulnerability-Scoring-System-AIVSS -how to measure/score AI and AI agents risks by Ken Huang

New Outline Proposal

"AI Autonomous Agents"

I. Executive Summary (1 page)

- **Purpose:** Brief overview of the importance, potential, and challenges of AI autonomous agents in cloud and enterprise contexts. The paper addresses identity and access concerns inherent to the distributed and automated nature of AI agents. Additionally, it will establish boundaries for secure design and provide industry-aligned examples to contextualize AI agent use in business settings, emphasizing the evolving role of human oversight in managing these systems.
- **Scope:** Summarize the paper's focus on security, governance, privacy, and practical implementation.The paper will define agentic systems, explaining their distinct characteristics compared to traditional systems through compare and contrast of their architecture non-functional attributes (trustability, explainability, non-deterministic nature and consumability) plus their data flows. Key areas of focus will include unique risks and necessary controls, with specific attention to data access, risk-mitigation techniques, system-to-system interactions, and inter-agent communication.

II. Introduction to ~~Autonomous~~ Multi-Agent System utilizing and Autonomy (1.5 pages)

AI agent systems, referred to in this paper as “a *multi-agent system* (MAS), is a system of AI agents working cooperatively interacting in a shared environment to achieved goals. MAS architecture is the default architecture and design explored. A MAS aligns on a common overarching output goal but within the MAS each AI agent likely has differing rulesets, rewards, and sub-goals.

Definition and Evolution: Define autonomous agents and contrast them with traditional automated systems. AI agent autonomy refers to the degree to which an artificial intelligence system or agent can make independent decisions, take actions, and adapt to new information without requiring human intervention. Autonomy in AI involves the following characteristics:

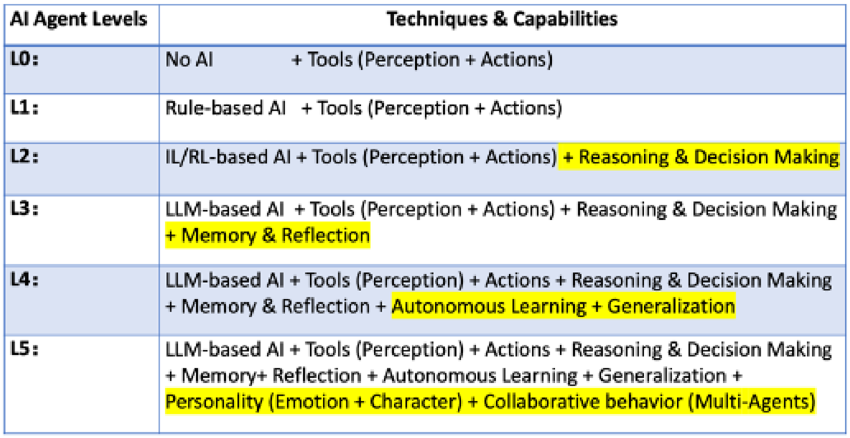
- **Goal orientation:** designed with a goal or objective in mind; directing actions toward achieving goals, rewards, or solution to a specific problem.
- **Perception:** senses and interpret its environment - typically through sensor input and data input
- **Decision-making**: selecting a decision best on based on given goals, given rules, or learning algorithms
- **Execution:** carry out tasks or actions

The aggregate impact of combining perception, decision-making, goal orientation and execution are components of *adaptability* - a non-functional software architectural requirement demonstrating learning from experience, adapting behavior to new circumstances, and improving performance. Adaptability is an element of Autonomy. Autonomy ranges from ‘low’ (agents requiring significant human guidance) to ‘high’ (agents functioning entirely independently), depending on the level of decision-making power and flexibility programmed into the AI system.

- **Components of Agentic Systems:**
  - Discuss common architectural **elements** (e.g. decision-making modules, perception, action mechanisms, and communication protocols)
  - Agentic classification **archetypes:** by business use case, by planning type (e.g., supply chain), by tool access, by multi-agent specialization (e.g., virtual assistant), by chain of thought (e.g. solving a math problem, kill chain action), by hierarchy (i.e., project management skill routing and task delegation)
  - Elaborate from simple agent **patterns,** simplified agentic system patterns, to orchestrated agents, to co-dependent agentic systems advancing along increasing levels of autonomy.
  - Rise of compound patterns: reflection + chain of thought; planning + multi-agent; tool use + memory augmentation, and so forth.
  - AI Agent Taxonomy (from AI Controls Framework Workgroup URL and URL)
- **Data Flow and Control Mechanisms:.**
  - Explain high-level data flows and how agents interact with other systems, databases, structured and unstructured data, and API’s (i.e., along the edge of the model adjacent to design layer).
  - Explain how tools are created, called and how the outputs are used
  - Traceability and explainability considerations as calls are made across systems
- **Inter-Agent Communication and Orchestration:**
  - Outline how agents communicate and coordinate within a system, introducing terms like "Human-in-the-Loop" (HITL) and "Fail-Safe" mechanisms for oversight.
    - (i.e., adjacent to the edges of the scope boundary)
      - Parallel autonomy (HITL): Share control based upon what the human wants to accomplish. Parallel autonomy prevents risky human behavior in the field of operation.
      - Series Autonomy (Failsafe/circuit breaker operation) This is contingent upon safe system operation throughout the operating environment.
    - Performance monitoring and model monitoring, placement and analysis of performance and model monitoring data (e.g., is it a feedback mechanism as well as an oversight mechanism? The monitoring data has two different use cases).
    - Adjustments over time as agent performance and reliability improves

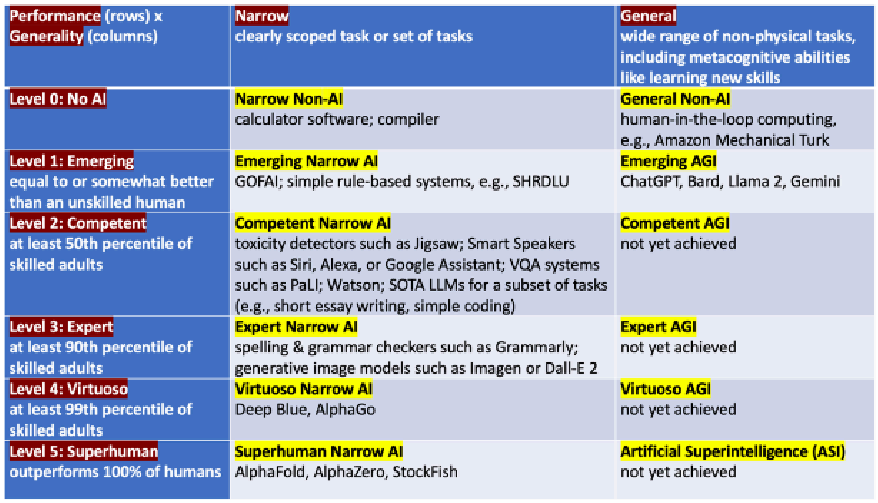
AI Agent Autonomy https://arxiv.org/pdf/2405.06643

Table 3. Levels of AI Agents



AI Agent Levels	Techniques & Capabilities
L0:	No AI + Tools (Perception + Actions)
L1:	Rule-based AI + Tools (Perception + Actions)
L2:	IL/RL-based AI + Tools (Perception + Actions) + Reasoning & Decision Making
L3:	LLM-based AI + Tools (Perception + Actions) + Reasoning & Decision Making + Memory & Reflection
L4:	LLM-based AI + Tools (Perception) + Actions + Reasoning & Decision Making + Memory & Reflection + Autonomous Learning + Generalization
L5:	LLM-based AI + Tools (Perception) + Actions + Reasoning & Decision Making + Memory+ Reflection + Autonomous Learning + Generalization + Personality (Emotion + Character) + Collaborative behavior (Multi-Agents)

Table 1. Levels of AGI [28]



Performance (rows) x Generality (columns)	Narrow clearly scoped task or set of tasks	General wide range of non-physical tasks, including metacognitive abilities like learning new skills
Level 0: No AI	Narrow Non-AI calculator software; compiler	General Non-AI human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging equal to or somewhat better than an unskilled human	Emerging Narrow AI GOFAI; simple rule-based systems, e.g., SHRDLU	Emerging AGI ChatGPT, Bard, Llama 2, Gemini
Level 2: Competent at least 50th percentile of skilled adults	Competent Narrow AI toxicity detectors such as Jigsaw; Smart Speakers such as Siri, Alexa, or Google Assistant; VQA systems such as PaLI; Watson; SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	Competent AGI not yet achieved
Level 3: Expert at least 90th percentile of skilled adults	Expert Narrow AI spelling & grammar checkers such as Grammarly; generative image models such as Imagen or Dall-E 2	Expert AGI not yet achieved
Level 4: Virtuoso at least 99th percentile of skilled adults	Virtuoso Narrow AI Deep Blue, AlphaGo	Virtuoso AGI not yet achieved
Level 5: Superhuman outperforms 100% of humans	Superhuman Narrow AI AlphaFold, AlphaZero, StockFish	Artificial Superintelligence (ASI) not yet achieved

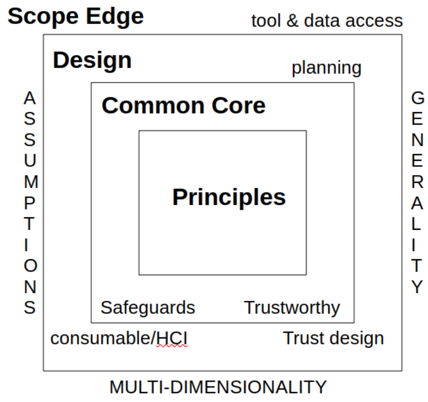
III. Core Architecture of Autonomous Agents (2 pages)

Gathered visual interpretation from 10/30/2014 meeting discussion

Multi-Agent Systems (MAS) represent a system of artificial intelligent agents who combine to deliver business value when aligned to a business objective or objectives. The default agentic system is multiple agents enjoined to accomplish a business goal.

MAS architectural patterns are,

1. **Singular:** multiple agents individually connected to a larger third party or internal model.
2. **Networked**: multiple agents connected to each other in a mesh arrangement
3. **Supervisory hierarchy**: One agent serves as a master controller for subordinate agents
4. **Supervised serialized**: agent to agent connectivity through input-output policy management control points.



-

IV. Security Implications for Autonomous Agents (2 pages)

- ~~**Threat Landscape:**~~
    - Identify the key security risks unique to AN autonomous agent, such as adversarial attacks, data poisoning, data privacy breaches, and software vulnerabilities.
    - Particular focus on indirect prompt injection as an attack vector
    - Suggestion to add Contextual Manipulation in addition to indirect prompt injection, consider discussing the potential for contextual manipulation in autonomous agents that rely heavily on large-scale models like LLMs or knowledge graphs (e.g., attackers could easily exploit the agent’s understanding of the world, shaping its behavior via subtle nudges in context or framing.)
- ~~**Attack Surfaces:**~~

    - Analyze potential entry points for attacks, including APIs, sensor inputs, communication protocols, and decision-making algorithms.
    - Assess alternative attack surfaces: introducing model error, introducing bias variance, auto-correlation manipulation, thinning the subset data, i.e., attack the app program, attack the (math) method, introduce an environmental change, attack the edges of a NN,
- ~~**Basic Defense Strategies:**~~
    - AI and ML defense is a team sport. Outline foundational measures like anomaly detection, encryption, and intrusion detection to mitigate security risks and how these will change due to emergent AI agentic capability ~~(i.e., agent vs agent combat - we see this now with HFT algorithms reference 1987 Wall Street Flash Crash).~~
        - ~~Advanced topic~~ (...what is old is new again) 'semantic attacks" and 'simula warfare' page 79 https://apps.dtic.mil/sti/tr/pdf/ADA367662.pdf.
        - ~~Danny Hillis outcomes:~~ What happens when AI acts on its own interests or multiple corporatized AI agents act aligned to the interests of their corporate masters? (e.g., the 1987/2010 algorithmic HFT flash crashes)

V. Risk Management and Governance (2 pages)

- **Risk Assessment Models:**
    - Introduce risk frameworks for assessing and managing autonomous agent risks (NIST, ISO standards, and other relevant frameworks).
    - *How traditional risk formulas need to be extensible or used in combination* to address facets of risk (e.g. ALE + FMEA; FMEA + FTA; Monte Carlo + RIE and so forth)
- **Regulatory Compliance and Industry Standards:**
    - Overview of regulations impacting autonomous agents (e.g., EU AI Act, U.S. privacy laws) and alignment with industry standards.
    - CONUS and OCONUS *Heat map* of most important regulations that will impact the deployment and use of AI agents, and agentic systems.
    - Climate and Energy impact of AI usage (Data Center needs, data center construction)
    - Operating environment: laws not harmonized, policies not harmonized, liability not harmonized, geographic boundaries eroded, controlling for contradictory behavior and self-contradictory behavior, who owns the decisions and the rewards system for AI agents.
- **Transparency and Accountability:**
- It is easier to assess a solution for fitness, than it is to design one**.**
- "Reality apathy" is when the human viewer or participant does not know what can be believed, which destroys trust.
    - Discuss the importance of transparency in decision-making processes, auditing, and accountability structures including climate impact.

  - https://ai.google/responsibility/responsible-ai-practices/
  - https://openai.com/index/openai-safety-update/ and https://openai.com/index/securing-research-infrastructure-for-advanced-ai/
  - https://carbonaccountingfinancials.com/ (is this reg. coming for every industry?)
  - https://hbr.org/2024/07/the-uneven-distribution-of-ais-environmental-impacts
  - https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about
  -

VI. Privacy and Data Protection Considerations (1.5 pages)

- **Data Access Controls:**
  - Examine access controls and permissions to ensure data privacy within autonomous systems.
- **Data Minimization and Anonymization Techniques:**
  - Traditional detail approaches like data anonymization, particularly for compliance with data protection standards (e.g., GDPR).
  - Data minimization is hard. (OOP data hiding, interface contracts, encapsulation, DRM, de-aggregate, mask/obfuscate methods, tokenize, redact, encryption, compression, de-identify, anonymize).
- What are the novel “data minify” approaches?
  - Differential Privacy, Federated Learning (train locally/share updated only), Synthetic data generation, Partial homomorphic encryption, ZKP (zero knowledge proofs, embedding, feature hashing, sensitive data GAN anonymizing, gaussian noise.
- **Interaction with Third-Party Systems:**
  - Explore the implications of autonomous agents interacting with external systems, emphasizing API security and third-party integration risks.

VII. Ethical Considerations (1 page)

- **Bias and Fairness in Decision-Making:**
  - Discuss the ethical challenges of bias in autonomous agents and ways to mitigate its impact.
- **Human Oversight and Control:**
  - Outline roles for human oversight in decision-making to prevent unintended actions or outcomes. (see series vs. parallel autonomy)
- **Accountability:**
  - Address the need for transparent and accountable agent systems in high-stakes applications.

VIII. Practical Implementation and Best Practices (2 pages)

- **Design Principles for Secure AI Agent Systems:**
  - Provide actionable design principles, focusing on scalability, security, and interoperability in cloud environments.
  - Emphasize practical strategies for reducing complexity and enhancing security, such as modular design, regular updates, and secure communication protocols.
  - Emergine Control Matricies (CSA 'AICM' URL)
- ~~**Integration with Existing IT Systems:**~~
  - ~~Guide on how to incorporate autonomous agents into existing IT infrastructures, including microservices architectures and data management.~~
- Insert Rajesh K. work on AI Agentic framework to assess maturation with respect to harm.
- Financial return on investment for AI Agentic Systems. (placeholder - this is an opportunity to develop a spreadsheet tool).

~~**IX. Use Cases in Industry (1.5 pages) - maybe omit?**~~

- ~~**Healthcare:**~~
  - ~~Discuss predictive maintenance of medical equipment and its impact on patient care.~~
- ~~**Transportation:**~~
  - ~~Explore anomaly detection in traffic systems and optimization of transport flow.~~
- ~~**Energy and Critical Infrastructure:**~~
  - ~~Examine autonomous agents for monitoring power grids a~~
  - ~~nd managing energy resources.~~

~~**X. Future Trends and Research Directions (1.5 pages) - depends on how the earlier content goes.**~~

- ~~**Advances in AI and Machine Learning Algorithms:**~~
  - ~~Overview of emerging techniques that could enhance the capabilities and resilience of autonomous agents.~~
- ~~**Regulatory Landscape: (duplicative?)**~~
  - ~~Discuss anticipated developments in global regulations and compliance frameworks for AI systems.~~
- ~~**Opportunities for Enhanced Security and Privacy: (duplicative too?)**~~
  - ~~Consider evolving methods for fortifying security and privacy, such as zero trust architectures and federated learning.~~

XI. Conclusion and Recommendations (0.5 page)

- **Key Takeaways:**
  - Recap the core findings, risks, and best practices.
- **Strategic Recommendations:**
  - Provide actionable steps for executives and engineers to prepare for and deploy AI autonomous agents effectively.

Draft of Splitting Components and methodologies to complement III. above

1. Components of Agentic Systems

Agentic systems are composed of various interconnected components, each contributing to the overall functionality of the system. These include:

Decision-Making Modules: These form the core of an agent's cognitive abilities, using algorithms like decision trees, neural networks, or reinforcement learning models to make informed decisions.

Perception Mechanisms: Agents perceive their environment through a variety of sensory inputs—ranging from vision and audio to specialized sensors (e.g., LIDAR for self-driving cars). These inputs are processed into actionable data that informs the agent's decisions.

Action Mechanisms: Once a decision is made, agents execute it through action mechanisms, which could involve controlling robots, sending commands to other software, or triggering external processes.

Communication Protocols: Autonomous agents communicate both internally with other agents and externally with humans or systems. Protocols range from APIs to low-level networking stacks or more complex multi-agent communication frameworks.

Memory: To adapt and plan over time, agents store information about past actions, decisions, and environmental states. This enables them to modify their behavior based on previous outcomes, improving their problem-solving abilities.

External Inputs and Outputs: Agents interact with external systems and data sources, ranging from web APIs to IoT devices. These inputs might include sensor readings, user instructions, or market data, while outputs could involve reports, control commands, or system status updates.

2. Methodologies for Problem-Solving in Agentic Systems

Once the components are in place, agents rely on a variety of methodologies to solve problems, plan tasks, and take action. Key approaches include:

Chain of Thought (CoT): Agents use CoT to break down complex tasks into smaller, manageable steps. This can be seen in applications like question-answering or problem-solving, where the agent follows a logical sequence of steps to reach a conclusion.

Reflection: Through reflection, agents continuously evaluate their past actions and adapt their behavior. This is crucial for systems that need to self-correct or learn from mistakes (e.g., reinforcement learning).

Tool Use: Agents often use external tools—whether libraries, services, or physical tools—to augment their capabilities. In a multi-agent system, tool usage might also involve collaborative use of shared resources or systems.

Memory Augmentation: Some agents have extended memory that allows them to store and retrieve important information over time, aiding in decision-making processes and making them more adaptable to dynamic environments.

Multi-Agent Coordination: In complex environments, agents may need to work together to accomplish a goal. Techniques like task delegation, cooperative planning, and negotiation ensure that agents function in harmony.

Hierarchical Task Decomposition: For large-scale problems, agents can decompose tasks hierarchically. This allows higher-level agents to assign sub-tasks to lower-level agents, ensuring efficient use of resources and better organization of efforts.

Lead Author Interest (feel free to add your name)

- Ed Sewell
- Devon Artis (Hands on experience with multiple agentic frameworks)
- Akram Sheriff (Detailed hands-on Project and Product First hand experience in LLM security Agents building)
- MJ Schwenger
- Manas Khanna
- Michael Roza

- Alex Rebo
- Aditya Garg
- Vaibhav Malik
- Craig Ellrod
- Rajesh Kanungo
- Pratixa Joshi
- Akshay Bhardwaj
- Govindaraj Palanisamy

- Keith Pasley
- Sudesh Gadewar
- ‘Lanre Bakare
- Pranav Kumar
- Sven Olensky
- Ivan Djordjevic (or contributor)
- Adam Lundqvist (rather contributor, though. Handson experience building ai agents for offensive security)
- Madhav Chablani
- Ashish Vashishtha
- Akash Mukherjee
- Manish Mishra
- John Hooks (Enterprise Hyperautomation AI Agency Implementation, Human-Robot Interaction (HRI) Research, Autonomous Humanoid Robotics.