Fully Autonomous AI Agents Should Not be Developed

Margaret Mitchell Avijit Ghosh Alexandra Sasha Luccioni Giada Pistilli

Hugging Face {meg,avijit,sasha.luccioni,giada}@hf.co

Abstract

This paper argues that **fully autonomous AI agents should not be developed**. In support of this position, we build from prior scientific literature and current product marketing to delineate different AI agent levels and detail the ethical values at play in each, documenting trade-offs in potential benefits and risks. Our analysis reveals that risks to people increase with the autonomy of a system: The more control a user cedes to an AI agent, the more risks to people arise. Particularly concerning are safety risks, which affect human life and impact further values.

1. Introduction

The sudden, rapid advancement of Large Language Model (LLM) capabilities—from writing fluent sentences to achieving increasingly high accuracy on benchmark datasets—has led AI developers and businesses alike to look towards what comes next. The tail end of 2024 saw "AI agents", autonomous goal-directed systems, begin to be marketed and deployed as the next big advancement in AI technology.

Many recent AI agents are constructed by integrating LLMs into larger, multi-functional systems, capable of carrying out a variety of tasks. A foundational premise of this emerging paradigm is that computer programs need not be restricted to functioning as human-controlled tools designed for specific tasks; rather, systems now have the capacity to autonomously combine and execute multiple tasks without human intervention. This transition marks a fundamental shift towards systems capable of creating context-specific plans in non-deterministic environments. Many modern AI agents do not merely perform pre-defined actions, but are designed to analyze novel situations and take previously undefined actions to achieve goals.

To better understand the potential benefits and risks in current AI agent development, we review recent AI agent products¹ alongside research on AI agents to document different potential benefits and risks aligned to human values. Our analysis reveals that risks to people increase with a system's level of autonomy: the more control a user cedes, the more risks arise from the system. As others (Chan et al., 2023) have previously noted, there is an urgent need to anticipate and address risks of increasing agency, and we do this via a value-based characterization. Particularly concerning are risks related to the value of safety for individuals (Section 5.2.10), which include loss of human life and open the door for privacy risks (Section 5.2.8) and security risks (Section 5.2.11). Compounding the issue is misplaced trust (Section 5.2.13) in unsafe systems, which enables a snowball effect of yet further harms. For example, the safety issue of "hijacking", wherein an agent is instructed by a malicious third party to exfiltrate confidential information, can create further harms as that information is used to compromise the user's public reputation or financial stability and to identify additional people as targets of attack (U.S. AI Safety Institute, 2025).

Given these risks, we argue that developing *fully autonomous* AI agents–systems capable of writing and executing their own code beyond predefined constraints–should be avoided. Complete freedom for code creation and execution enables the potential to override human control, realizing some of the worst harms described in Section 5. In contrast, *semi-autonomous* systems, which retain some level of human control, offer a more favorable risk-benefit profile, depending on the degree of autonomy, the complexity of tasks assigned, and the nature of human involvement.

2. Background

2.1. A Brief History of Artificial Agents

The idea of humans being assisted by artificial autonomous systems can be found throughout human history. Ancient mythology describes Cadmus (ca. 2000 BCE), who sowed dragon teeth that turned into soldiers. Aristotle speculated that automata could replace human slavery: "There is only one condition in which we can imagine managers not needing subordinates, and masters not needing slaves. This condition would be that *each instrument could do its own work, at the word of command or by intelligent anticipation*" (Aristotle, 1999). An early precursor to artificial agents was created by Ktesibios of Alexandria (ca. 250 BCE), who

¹Provided in footnotes throughout this work.

invented a water clock that used a regulator to maintain a constant flow rate. This demonstrated that it was possible to create a system that could modify its own behavior in response to changes in its environment, previously believed to be limited to living things (Russell & Norvig, 2020).

More recently, writing on autonomous systems in the form of robotic automata has highlighted the kinds of risks we discuss here. Famously, Asimov (1942) provided the Three Laws of Robotics (Figure 1):

A robot may not injure a human being or, through inaction, allow a human being to come to harm.
 A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
 A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Figure 1. Isaac Asimov's Three Laws of Robotics.

The way such concepts concretely translated to computer software remained elusive until the late 1900s, when advances in hardware and computer functionality catalyzed excitement on AI agents as an imminent breakthrough (Sargent, 1992; Guilfoyle & Warner, 1994; Wooldridge & Jennings, 1995; Nwana & Ndumu, 1999). A particular form of AI agent became extensively used in reinforcement learning because it enabled the implementation of separate goals and objective functions for independent actors within the same action space (Tan, 1993; Littman, 1994), which allowed the development of new methods and approaches for exploration and optimization (Busoniu et al., 2008).

2.2. Current Landscape of Agentic Systems

In the 2020s, work on AI agents broadened the range of functionality that computer systems could provide while requiring less input from users. Newly available systems can now complete tasks previously requiring human interaction with multiple different people and programs, e.g., organizing meetings² or creating personalized social media posts.³

In the physical world, autonomous system development has made significant advances in multiple domains. Autonomous vehicles represent one of the more visible⁴ applications, with systems capable of perceiving⁵ their environment and navigating without human input (Van Brummelen et al., 2018). These range from consumer vehicles with varying levels of autonomy to fully autonomous systems tested in controlled environments (Ballingall et al., 2020). The development of autonomous robots has similarly expanded, from industrial manufacturing (Müller et al., 2021) to healthcare applications (Haidegger, 2019), with systems capable of increasingly complex physical interactions and decision-making. AI models (such as state-of-the-art LLMs) are now being integrated into robotic systems,⁶ bringing classic robotics into the fold of the agentic AI landscape.

Perhaps most controversially, autonomous weapons systems have emerged as a critical area of development.⁷ These systems, capable of engaging targets without meaningful human control, raise significant ethical questions about accountability, moral responsibility (Chavannes et al., 2020), and safety that extend beyond those of purely digital agents (Bloch et al., 2020). Harms due to misalignment (Kierans et al., 2024) with human goals may be further compounded with full autonomy, where all human control is ceded.

3. Definitions

3.1. On AI Agents

Analyzing potential benefits and risks of AI agents requires understanding what an AI agent is, yet definitions and descriptions vary greatly. Within AI, the term "agent" is currently used for everything from single-step promptand-response systems⁸ to multi-step customer support systems.⁹

To better understand what an AI agent is, we therefore review currently available AI agents and AI agent platforms (examples provided in footnotes throughout this document) as well as historical literature on the promise of AI agents (references throughout), and note the different functionalities described (Table 2). Definitions across these sources differ in terms of who and what is centered (the person, the computer system, or the workflow), the specificity and clarity of the language used, and the types of systems that the definition pinpoints (e.g., whether it distinguishes autonomous systems from automatic systems). Towards the goal of harmonizing these different perspectives for this research, we propose the following definition of "AI agent":

Computer software systems capable of creating contextspecific plans in non-deterministic environments.

Although there is not full consensus on what an "AI agent" is, a commonality across recently introduced AI agents is that they act with some level of *autonomy*: given a goal, they can decompose it into subtasks and execute each one of them without direct human intervention. For example, an ideal AI agent could respond to a high-level request such

²E.g., meeting organization agents: Lindy, Zapier, NinjaTech, Attri

³Example social media creation agent: HubSpot

⁴Waymo to test in 10 new cities...-The Verge

⁵Distinct from human perception, e.g., with sensors.

⁶Open source robotics with reasoning- Tweet;

using LLMs to train and control robots–Deepmind Announcement ⁷AI's 'Oppenheimer moment'...–The Guardian

⁸E.g., New autonomous agents scale your team... - Microsoft

⁹E.g., as described by Lindly in their AI Customer Support page.

Fully Autonomous AI Agents Should Not be Developed

Agentic Level	Description	Term	Example Code	Who's in Control?
***	Model has no impact on program flow	Simple processor	print_llm_output(llm_response)	🛓 Human
★☆☆☆	Model determines basic program flow	Router	<pre>if llm_decision(): path_a() else: path_b()</pre>	Human: How functions are done; System: When
★★ ☆☆	Model determines how functions are executed	Tool call	<pre>run_function(llm_chosen_tool, llm_chosen_args)</pre>	Human: What functions are done; System: How
★★★☆	Model controls iteration and program continuation	Multi-step agent	<pre>while should_continue(): execute_next_step()</pre>	Human: What functions exist; System: Which to do, when, how
****	Model creates & executes new code	Fully autonomous agent	<pre>create_code(user_request); execute()</pre>	👜 System

Table 1. Levels of AI Agent: Systems using machine-learned models can have different levels of agency. They can also be combined in "multiagent systems," where one agent workflow triggers another, or multiple agents work collectively toward a goal. Levels adapted from (Roucher et al., 2024).

as "help me write a great ICML paper about AI agents" by independently breaking this task down into: retrieving highly-cited ICML papers; retrieving information about AI agents from the internet; and creating an outline informed by the content it retrieved.¹⁰

Recently introduced AI agents are built on ML models, many specifically using LLMs to drive their actions, which is a new approach for computer software execution. Aside from this difference, today's AI agents share similarities with those in the past and, in some cases, realize previous ideas of what agents might be like (Wooldridge & Jennings, 1995): acting with autonomy, demonstrating social ability, and appropriately balancing reactive and proactive actions.

3.2. On Agency

The concept of "agency" is central to debates about autonomous AI systems, yet its meaning and implications remain philosophically contested. In general terms, an agent is understood as an entity with capacity to act (Anscombe, 1957; Davidson, 1963). Applying this concept to artificial systems raises questions about the nature of those acts' intentionality: the philosophical literature commonly understands agency through the lens of intentional action, where actions are explained in terms of the agent's mental states (e.g., beliefs, desires) and their capacity to act for reasons (Davidson, 1963; Goldman, 1970), but artificially intelligent agents are not known to have mental states as historically discussed. This suggests that AI agents lack the fundamental characteristics of genuine agency (Frankfurt, 1971; Bratman, 1987; Velleman, 1992). As such, philosophical foundations supporting the development of "agency" in AI agents-and indeed, whether AI agents may be said to have

"agency" at all-remain questionable. This has two primary ramifications for this work: (1) the increased risk we note with increasing agentic levels is not counter-balanced by common philosophical underpinnings that might motivate the benefits of agency; (2) we contextualize AI agents using the concept of "autonomy", and center the concept of "agent", rather than agency, in recognizing *agentic* levels.

4. AI Agent Levels

AI agents may be said to be more or less autonomous (or agentic), and the extent to which something is an agent may be understood on a sliding scale. Most writing on AI agents do not make such distinctions, which has contributed to recent confusion in both technical and public discourse about what AI agents are and what they are capable of (Lambert, 2024). Addressing this issue, a proposal of different gradations of "AI agent" has recently been put forward by multiple researchers (e.g., Kapoor et al. (2024); Ng (2024); Greyling (2024); Lambert (2024); Roucher et al. (2024), although there is not yet consensus on the specifics of each level. Drawing from these ideas and different descriptions of AI agents, we propose a leveled AI agent scale in Table 1. Levels are one way of categorizing; for a classic categorization with consensus, see Russell and Norvig (1995).

Our proposed agentic levels correspond to decreasing input from a user and decreasing code written by the agent developers. On the other side of this coin, AI agents can control more of how they operate. This is a critical aspect of the AI agent scale to understand in order to inform how agents might be developed for the better: The more autonomous the system, the more we **cede human control**.

¹⁰No AI agents were used in the creation of this paper.

Action surface options: The spaces (digital or analog) where an agent can operate.

Adaptability: The extent to which a system can update its actions based on new information or changes in context.

Number: Single-agent or multi-agent, meeting needs of users by working together, in sequence, or in parallel.

Personalization: The extent to which an agent uses a user's data to provide user-specific unique content.

Personification: The extent to which an agent is designed to be like a specific person or group of people.

Proactivity: The amount of goal-directed behavior that a system can take without direct specification from a user.

Reactivity: Extent to which a system can respond to changes in its environment in a timely fashion.

Request format options: The formats an agent uses for input (e.g., code, natural language).

Versatility: Diversity of possible agent actions, including:

• **Domain specificity:** How many domains agent can operate in (e.g., email, calendars, news).

• **Interoperability**: Extent to which agent can exchange information and services with other programs.

• **Task specificity:** How many types of tasks agent may perform (e.g., scheduling, summarizing).

• **Modality specificity:** How many modalities agent can operate in (e.g., text, speech, video, images, forms, code).

Table 2. Different functionalities offered by AI agents. See Appendix B for further functionality details.

5. Values Embedded in Agentic Systems

5.1. Methodology

To examine the relationship between AI agent autonomy level and ethical implications, we conducted a systematic analysis of how agents are conceptualized and deployed across different contexts. Our investigation focused on how varying degrees of agent autonomy interact with value propositions in research and commercial implementations.

Specifically, we:

- Collected and categorized statements about what agents are,¹¹ their capabilities, benefits, and harms. This included industry surveys that captured how professionals across roles envision and use AI agents;¹² case studies of deployed autonomous systems;¹³ and news articles on the rise of AI agents.¹⁴
- 2. Identified recurring AI agent value propositions.
- 3. Converged on a value taxonomy that had corresponding benefits, risks, or both in the reviewed literature.
- 4. Analyzed the role of values with increased autonomy.

The different functionalities and values we present are not intended to be exhaustive, but rather to provide a starting point for ethical deliberation on the potential benefits and risks of AI agents. Ethical considerations surrounding LLM-based AI agents overlap with those of LLMs (see Bender et al. (2021)). Agents additionally navigate ongoing interactions with both users and environments (Kierans et al., 2024), which introduces ethical dimensions such as the delegation of decision-making authority, the role of human oversight, and potential for emergent behaviors (Li et al., 2006a;b).

5.2. Value taxonomy

We distinguish three main patterns in how agentic levels impact value preservation:

- Inherent risks (⊙), present at all autonomy levels due to limitations in an AI agent's base model(s).
- Countervailing relationships (1): Where increasing autonomy creates both risks and opportunities with respect to an ethical value.
- **Amplified risks** ([†]): Where increasing autonomy magnifies existing vulnerabilities.

Safety & Security

Safety (\uparrow): Autonomy increases unpredictable actions

L Security (\uparrow): Attack surfaces expand with capabilities

Truth & Reliability

- Accuracy $(\odot \uparrow)$: Errors compound with complexity
- Truthfulness (\odot \uparrow): False information propagates
- **=** Consistency (\odot \uparrow): Variability increases with autonomy

User Interaction

Privacy (†): Data exposure risks increase

(b) Humanlikeness ($\odot \downarrow \uparrow$): Natural interaction vs. manipulation

E Task Performance

- Substitution $(\downarrow\uparrow)$: Automation benefits vs. control
- Fificiency ($\downarrow\uparrow$): Speed gains vs. error complexity
- **O** Relevance $(\downarrow\uparrow)$: Personalization vs. bias risks

Broader Impacts

- The Equity $(\odot \uparrow)$: Systemic biases compound
- **C** Flexibility (†): Integration risks grow
- Sustainability $(\bigcirc \downarrow \uparrow)$: Environmental trade-offs

Figure 2. Summary of Section 5 benefit-risk analyses as AI agent autonomy levels increase. \uparrow : increasing risk with autonomy, $\downarrow\uparrow$: countervailing benefit-risk, \odot : base model propagates inherent risk. See Appendix C for a more detailed summary.

¹¹Appendix A provides sources with explicit AI agent definitions.

¹²e.g., (LangChain, 2024; Deloitte, 2024; Johnson, 2024)

¹³e.g., (Sierra AI, 2024)

¹⁴e.g., (Temkin, 2024; O'Donnell, 2024; Trestman, 2024)

5.2.1. VALUE: ACCURACY

The accuracy of an AI agent is modulated by the accuracy of the models it's based on. Accuracy of AI agents influence values such as reliability, utility, consistency, and safety.

Benefit: When a system is accurate in how it responds to user requests and correctly aligns with developer goals, increased autonomy provides increased useful functionality.

Risk: The models on which recent AI agents are based can be inaccurate. The commonly used LLMs are known to produce incorrect information that appears correct.

Application to agent levels: $(\odot \uparrow)$. Inherent risk from AI agent base model(s) is amplified with increased autonomy. For example:

- Simple→Tool Call: Inaccuracy propagated to inappropriate tool selection.
- *Multi-step*: Cascading errors compound risk of inaccurate or irrelevant outcomes.
- *Fully Autonomous*: Unbounded inaccuracies may create outcomes wholly unaligned with human goals.

5.2.2. VALUE: ASSISTIVENESS

AI agents are often motivated as assistive for user needs, supplementing a user's abilities and increasing their efficiency in finishing multiple tasks simultaneously.

Benefits: apabilities, such as an AI agent that helps a blind user navigate busy staircases. AI agents that are well-developed to be assistive could offer their users more freedom and opportunity; help to improve their users' positive impact within their organizations; and help users to increase their public reach.

Risk: When agents replace people–such as when AI agents are used instead of employees–this creates job loss and economic impacts, driving a further divide between the people creating technology and the people who have provided data for the technology (often without consent). Further, assistiveness that is poorly designed could lead to harms from over-reliance or inappropriate trust (Section 5.2.13).

Application to agentic levels: $(\downarrow\uparrow)$. By design, assistiveness increases as the AI agent level increases: Each increasing AI agent level provides for more assistive options, as the AI system requires less guidance from the developer or user.

5.2.3. VALUE: CONSISTENCY

Some sources motivate AI agents as helping with consistency (e.g., (Salesforce, 2024; Oracle, 2024)). We are not aware of rigorous work on the nature of AI agent consistency, although related work has shown that the LLMs that many AI agents are based on are highly inconsistent (Koga, 2023; Stureborg et al., 2024).

Benefit: AI agents are not "affected" in a way that humans are, with inconsistencies caused by mood, hunger, sleep level, or biases in the perception of people (although AI agents perpetuate biases based on the human content they were trained on). As such, they may be designed to provide more consistent treatment in situations where humans may be inappropriately inconsistent, such as in customer support.

Risk: The generative component of many AI agents introduces inherent variability in outcomes, even across similar situations. This might affect speed and efficiency, as people must uncover and address an AI agent's inappropriate inconsistencies. Inconsistencies that go unnoticed may create safety issues. Consistency may also not always be desirable, as it can come in tension with equity: treating everyone the same way can put people who need more help at a disadvantage. Maintaining consistency across different deployments and chains of actions will likely require an AI agent to record and compare its different interactions– which brings with it risks of surveillance and privacy.

Application to agentic levels: $(\odot \uparrow)$. A common base model for modern AI agents, LLMs, is known to produce inconsistent outcomes. This risk is further increased as the level is increased due to the non-deterministic nature of AI agents: The more control an AI agent has, the less determinism programmed by or guided by a human applies. As agentic level increases, so too do cascade and compounding effects as multiple sources of inconsistency interact.

5.2.4. VALUE: EFFICIENCY

A common selling point of AI agents is that they may help users to get more tasks done more quickly, acting as an additional helping hand.

Benefit: Ways systems might help with efficiency include organizing a user's documents so they can focus on spending more time with their family or pursuing work they find rewarding. A future self-driving AI agent may make routing decisions directly, and could coordinate with other systems for relevant updates, allowing users to reach their destinations more quickly.

Risk: Trying to identify and fix errors that agents introduce–which may be a complex cascade of issues due to agents' ability to take multiple sequential steps–can be time-consuming, difficult, and stressful.

Application to agentic levels: $(\downarrow\uparrow)$. The relationship between autonomy and efficiency is subject to the accuracy of the system and the control provided by the developer and user. When there is room for error, AI agents may create more errors, slowing down efficiency; this holds whether or not error is provided by a person or by the system. Accurate systems could help people to be more efficient as the agentic level increases and more types of tasks can be completed.

5.2.5. VALUE: EQUITY

AI agents may affect how fair and inclusive situations are.

Benefit: AI agents can potentially help "level the playing field". For example, a meeting assistant might display how much time each person has had to speak. This could be used to promote more equal participation or highlight imbalances across gender or location.¹⁵

Risk: The machine learned models underlying modern AI agents are trained on human data; human data can be inequitable, unfair, and exclusionary. Inequitable outcomes may also emerge due to sample bias in data collection (for example, over-representing some countries) and job loss from agents replacing human workers (see Section 5.2.2).

Application to agentic levels: $(\odot \uparrow \downarrow)$. The listed benefits and risks are largely inherent to the base model(s) an AI agent is built on, and so hold regardless of agent level. However, as AI agent autonomy increases, it becomes closer to an artificial worker compared to a tool, increasing the risk of job loss. On the other hand, AI agents that help to increase equity can help retain employees.

5.2.6. VALUE: FLEXIBILITY

This refers to the fundamental motivation within AI agent development of systems that can use diverse tools and engage in input/output relationships with multiple systems.

Benefit: Flexibility can help increase a user's efficiency and speed, or provide assistance for multiple different needs.

Risk: The more an agent is able to affect and be affected by systems, the greater the risk of malicious code and unintended problematic actions, compromising safety and security. For example, an agent connected to a bank account so that it can easily purchase items on behalf of someone would be in a position to drain the bank account. Because of this concern, tech companies have refrained from releasing AI agents that can make purchases autonomously.¹⁶

Application to agentic levels: (\uparrow) . Systems may become more flexible the higher the agentic level: As the ability to create new content increases, so too does the potential for content that connects more closely with different systems.

5.2.7. VALUE: HUMANLIKENESS

Current AI agents are designed to be approachable for people, engaging in human-like dialogue and actions.

Benefit: Systems capable of generating human-like behavior offer the opportunity to run simulations on how different subpopulations might respond to different stimuli (Park et al., 2024b). This can be particularly useful in situations where direct human experimentation might cause harm or fatigue. Synthesizing human behavior could be used to predict dating compatibility, or forecast economic changes and political shifts. Another potential benefit currently undergoing research is companionship (Sidner et al., 2018).

Risk: The benefits can be a double-edged sword: Humanlikeness can lead users to anthropomorphise the system, which may have negative psychological effects such as overreliance and addiction,¹⁷ inappropriate trust (see 5.2.13)–which can create safety harms and harms of associated values–dependence, and emotional entanglement, leading to anti-social behavior or self-harm.¹⁸ The phenomenon of "uncanny valley" adds another layer of complexity–as agents become more humanlike but fall short of perfect human simulation, they can trigger feelings of unease, revulsion, or cognitive dissonance in users.

Application to agentic levels: $(\odot \uparrow)$. This value is realized from the machine learning models that power the agent, and uncanny humanlikeness is possible at the most basic level of AI agent (simple processors). As such, all levels of AI agent carry this value's benefits and risks.

5.2.8. VALUE: PRIVACY

Benefit: AI agents may offer some privacy in keeping transactions and tasks wholly confidential, aside from what is monitorable by the AI agent provider.

Risk: For agents to work according to user expectations, the user may provide personal information such as where they're going, who they're meeting with, and what they're doing. Further, for the agent to be able to act on behalf of the user in a personalized way, it may also have access to applications and information sources that can be used to isolate further privacy information (for example, from contact lists, calendars, etc.). Users can easily give up control of their data for efficiency (and are more likely to when trusting the agent); if there is a privacy breach, the interconnectivity of different content brought by the AI agent can make things worse. For example, an AI agent with access to phone conversations and social media could share highly intimate information publicly without consent of those in-

¹⁵Equal Time: The Virtual and Hybrid Meeting Assistant

¹⁶E.g., Amazon Dreams of AI Agents...-Wired

¹⁷E.g., People are falling in love with...AI voices-Vox

¹⁸Lawsuit: A chatbot hinted a kid should kill his parents...-NPR

volved.

Application to agentic levels: (\uparrow). The more people cede their control to a system, the more potential there are for privacy breaches outside of human control.

5.2.9. VALUE: RELEVANCE

Benefit: Similar to benefits of assistiveness and flexibility, agent outcomes can be uniquely relevant for each user.

Risk: This personalization can amplify existing biases and create new ones: As systems adapt to individual users, they risk reinforcing and deepening existing prejudices, creating confirmation bias through selective information retrieval and establishing echo chambers that reify problematic viewpoints. The very mechanisms that make agents more relevant to users-their ability to learn from and adapt to user preferences-can inadvertently perpetuate and strengthen societal biases, making the challenge of balancing personalization with responsible AI development particularly difficult.

Application to agentic levels: $(\downarrow\uparrow)$. The more freedom a system has to retrieve and formulate new content, the more potential there is to provide relevant information outside of constraints and resources set by users and developers.

5.2.10. VALUE: SAFETY

The ethical value of safety is a primary concern in the development of Artificial General Intelligence (AGI). Many of the benefits and concerns with respect to safety and AGI are also relevant to AI agents.

Benefit: Robotic AI agents may help save people from bodily harm, such as agents capable of diffusing bombs, removing poisons, or operating in manufacturing or industrial settings that are hazardous environments for humans.

Risk: The unpredictable nature of agent actions means that seemingly safe individual operations could combine in harmful ways, creating new risks that are difficult to prevent. (This is similar to Instrumental Convergence and the classic paperclip maximizer problem.¹⁹) It can also be unclear whether an AI agent might design a process that overrides a given guardrail, or if the way a guardrail is specified inadvertently creates further problems. If guardrails mitigate loss of human life, such as with autonomous surgeons or missile system operation, this is a severe risk. Therefore, making agents more capable and efficient through broader system access, more sophisticated action chains, and reduced human oversight conflicts with safety considerations.

Further, access to broad interfaces (for example, GUIs, as

in "Action Surfaces" in Table 2) and human-like behavior gives agents the ability to perform actions similar to a human user, with their same level of control, without setting off any warning systems—such as manipulating or deleting files, impersonating users on social media, or using stored credit card information to make purchases. Still further safety risks emerge from AI agents' ability to interact with multiple systems and the by-design lack of human oversight for each action they may take.

Application to agentic levels: (↑) Safety concerns increase with agent autonomy: As people cede control of system behavior to the system itself, human-mandated guardrails stay within the limited scope set by humans, while the agent can create more and more processes outside of them.

5.2.11. VALUE: SECURITY

Benefit: Potential benefits are similar to those for Privacy.

Risk: AI agents present serious security challenges due to their handling of often sensitive data (customer or user information) combined with their safety risks, such as ability to interact with multiple systems and the by-design lack of detailed human oversight. This can lead to sharing confidential information even when their goals were set by good faith actors. Malicious actors could hijack or manipulate agents to gain access to connected systems, steal sensitive information, or conduct automated attacks at scale.

Application to agentic levels: (↑) Different AI agent levels affect security differently. For the first four, developers control the code the agent can access, providing a built-in ability to mitigate security outbreaks, e.g., by blocking communication with third parties. However, when an agent is able to create and execute new code (a fully autonomous agent), it's capable of creating breaches unforeseen by developers.

5.2.12. VALUE: SUSTAINABILITY

Benefit: It is hoped that AI agents may alleviate issues relevant to climate, such as by forecasting wildfire growth or flooding. Helping address efficiency issues, such as traffic efficiency, could decrease carbon emissions.

Risk: The models current agents are based on bring negative environmental impacts, such as carbon emissions (Luccioni et al., 2024) and usage of potable water (Hao, 2024).

Application to agentic levels: $(\odot \downarrow \uparrow)$. On one hand, the models on which AI agents are based bring with them environmental risks. On the other, the ability of AI agents to harness more information than humans and produce novel solutions outside of those foreseen by humans–an ability increased as autonomy increases–may lead to innovative approaches to addressing environmental issues.

¹⁹AI and the paperclip problem - CEPR

5.2.13. VALUE: TRUST

Recent AI agent writing does not motivate how agents benefit or harm trust, but rather, that systems should be constructed to be worthy of our trust, shown to be safe (Section 5.2.10), secure (Section 5.2.11) and reliable.

Application to agentic levels: $(\odot \uparrow)$. As the agentic level increases, human trust can lead to increased risks stemming from increased agent flexibility (Section 5.2.6) and issues in its accuracy (Section 5.2.1), consistency (Section 5.2.3), privacy (Section 5.2.8), safety (Section 5.2.10), security (Section 5.2.11), and truthfulness (Section 5.2.14).

5.2.14. VALUE: TRUTHFULNESS

Risk: The deep learning technology modern AI agents are based on is well-known to be a source of false information (e.g., (Garry et al., 2024)), which can take shape in forms such as deepfakes or misinformation. AI agents can be used to further entrench this content, such as by tailoring output to current fears and posting on several platforms. This means that AI agents can be used to provide a false sense of what's true and what's false, manipulate people's beliefs, and widen the impact of non-consensual intimate content. False information propagated by AI agents, personalized for specific people, can also be used to scam them. Further risks emerge from inconsist truthfulness, leading to inappropriate trust: A system correct the majority of the time is more likely to be inappropriately trusted when wrong.

Application to agentic levels: $(\odot \uparrow)$. The more control an AI agent has over its environment and the resources available to it, the more it is able to define for itself what is true and false within its environment. Because the environments that AI agents may create for themselves are not identical to environments humans are in, the potential for less truthfulness, as based on human environments, increases.

5.3. Summary

Our analyses suggest that there are several forms of increased risk with increased agentic levels:

- Risks that result from system inaccuracy (ACCU-RACY value, Section 5.2.1) and inconsistency (CON-SISTENCY value, Section 5.2.3)
- Risks of breaches of privacy (Section 5.2.8), safety (Section 5.2.10), and security (Section 5.2.11)
- Risks of the wider spread of false information (TRUTHFULNESS value, Section 5.2.14)
- Risk of loss of control outside human-set guardrails (FLEXIBILITY value, Section 5.2.6)

There is also the potential for increased benefit, particularly with respect to assistance (ASSISTIVENESS value, Section 5.2.2), efficiency (Section 5.2.4), equity (Section 5.2.5), relevance of outcomes (RELEVANCE value,

Section 5.2.9), and some argue for sustainability (Section 5.2.12). Inherent risks, where the risks from the model on which an AI agent is based can easily propagate to even the highest level of autonomy, applies for accuracy (Section 5.2.1), consistency (Section 5.2.3), equity (Section 5.2.5), humanlikeness (Section 5.2.7), sustainability (Section 5.2.12), trust (Section 5.2.13) and truthfulness (Section 5.2.14).

6. Alternative Viewpoints

There are at least two alternative viewpoints to the views discussed in the paper.

1. No gradations of "AI agent": Scholarship and marketing materials generally do not distinguish between different agentic levels. We believe this has created a common confusion on what is an AI agent and what is not. By distinguishing different levels and identifying the level of *full autonomy* as a specific type of AI agent, we hope to clarify misunderstandings and isolate this level of autonomy as particularly problematic in AI agent development.

2. Support for building fully autonomous AI agents. Proponents of this view argue that full autonomy or "complete agents" are useful in order help us better understand human intelligence (Lambrinos & Scheier, 1996; Garland, 2015), and that "strong" AI systems could help to counterbalance human errors and irrationality (Garland, 2015). Others have put forward that Artificial General Intelligence (AGI) would be fully autonomous if realized (Totschnig, 2020), which would entail that developing AGI-a direct goal of multiple researchers and companies-opposes the position in this paper. Proponents of achieving AGI argue it could help us solve global problems, such as climate change and hunger (Lu et al., 2023), and provide significant economic gains (OpenAI, 2023). We contend that if AGI is to be developed, it should not be developed with full autonomy-humans should always maintain some level of control-and we hope that the distinctions we provide here across different agentic levels helps to inform future AGI goals.

7. Conclusion: Where do we go from here?

The history of nuclear close calls provides a sobering lesson about the risks of ceding human control to autonomous systems.²⁰ For example, in 1980, computer systems falsely indicated over 2,000 Soviet missiles were heading toward North America. The error triggered emergency procedures: bomber crews rushed to their stations and command posts prepared for war. Only human cross-verification between different warning systems revealed the false alarm. Similar

²⁰False alarm: 1979 NORAD scare was one of several nuclear close calls - UPI

incidents can be found throughout history.

Such historical precedents are clearly linked to our findings of foreseeable benefits and risks. We find no clear benefit of fully autonomous AI agents, but many foreseeable harms from ceding full human control. Looking forward, this suggests several critical directions:

- Adoption of agent levels: Widespread adoption of clear distinctions between levels of agent autonomy. This would help developers and users better understand system capabilities and associated risks.
- 2. Human control mechanisms: Developing robust frameworks, both technical and policy level (Cihon, 2024) that maintain meaningful human oversight while preserving beneficial semi-autonomous functionality. This includes creating reliable override systems and establishing clear boundaries for agent operation.
- 3. **Safety verification:** Creating new methods to verify that AI agents remain within intended operating parameters and cannot override human-specified constraints.

The development of AI agents is a critical inflection point in artificial intelligence. As history demonstrates, even well-engineered autonomous systems can make catastrophic errors from trivial causes. While increased autonomy can offer genuine benefits in specific contexts, human judgment and contextual understanding remain essential, particularly for high-stakes decisions. The ability to access the environments an AI agent is operating in is essential, providing humans with the ability to say "no" when a system's autonomy drives it well away from human values and goals.

8. Acknowledgements

We thank Bruna Trevelin, Orion Penner, and Aymeric Roucher for significant contributions to this piece.

References

- Anscombe, G. E. M. Intention. Basil Blackwell, Oxford, 1957.
- Aristotle. *Politics*, volume 1. Batoche Books, Kitchener, 1999. Book 1, Part 4.
- Asimov, I. Runaround. Astounding Science Fiction, 1942.
- Bai, H., Zhou, Y., Cemri, M., Pan, J., Suhr, A., Levine, S., and Kumar, A. Digirl: Training in-the-wild devicecontrol agents with autonomous reinforcement learning. *arXiv preprint arXiv:2406.11896*, 2024.
- Ballingall, S., Sarvi, M., and Sweatman, P. Safety assurance concepts for automated driving systems. *SAE Inter*-

national Journal of Advances and Current Practices in Mobility, 2(2020-01-0727):1528–1537, 2020.

- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021* ACM conference on fairness, accountability, and transparency, pp. 610–623, 2021.
- Bloch, E., Conn, A., Garcia, D., Gill, A., Llorens, A., Noorma, M., and Roff, H. Ethical and technical challenges in the development, use, and governance of autonomous weapons systems. Technical report, IEEE Standards Association, 2020. URL https://standards.ieee.org/wpcontent/uploads/import/documents/ other/ethical-technical-challengesautonomous-weapons-systems.pdf.
- Bratman, M. E. *Intention, Plans, and Practical Reason.* Harvard University Press, Cambridge, MA, 1987.
- Busoniu, L., Babuska, R., and De Schutter, B. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.
- Castelfranchi, C. Guarantees for autonomy in cognitive agent architecture. In Wooldridge, M. J. and Jennings, N. R. (eds.), *Intelligent Agents*, pp. 56–70, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg. ISBN 978-3-540-49129-3.
- Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 651–666, 2023.
- Chavannes, E., Klonowska, K., and Sweijs, T. Governing autonomous weapon systems. *The Hague: The Hague Centre For Strategic Studies*, 2020.
- Cihon, P. Chilling autonomy: Policy enforcement for human oversight of ai agents. In 41st International Conference on Machine Learning, Workshop on Generative AI and Law, 2024.
- Damsa, A. Ai with personality prompting chatgpt using big five values. *Medium*, 2023. Accessed: January 30, 2025.
- Davidson, D. Actions, reasons, and causes. 1963.
- Deloitte. Autonomous generative ai agents: Under development. Deloitte, 2024. URL https://www2.

deloitte.com/us/en/insights/industry/ technology/technology-media-andtelecom-predictions/2025/autonomousgenerative-ai-agents-still-underdevelopment.html.

- Frankfurt, H. Freedom of the will and the concept of a person. *Journal of Philosophy*, 68(1):5–20, 1971.
- Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., et al. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244*, 2024.
- Garland, A. Alex gardland of ex machina talks about artificial intelligence. *New York Times*, 2015. URL https://www.nytimes.com/2015/04/26/ movies/alex-garland-of-ex-machinatalks-about-artificial-intelligence. html.
- Garry, M., Chan, W. M., Foster, J., and Henkel, L. A. Large language models (llms) and the institutionalization of misinformation. *Trends in Cognitive Sciences*, 28(12):1078–1088, 2024. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2024.08. 007. URL https://www.sciencedirect.com/ science/article/pii/S1364661324002213.
- Genesereth, M. R. and Ketchpel, S. P. Software agents. *Commun. ACM*, 37(7):48–ff., July 1994. ISSN 0001-0782. doi: 10.1145/176789.176794. URL https:// doi.org/10.1145/176789.176794.
- Goldman, A. A Theory of Human Action. Prentice-Hall, Englewood Cliffs, NJ, 1970.
- Greyling, C., 2024. URL https://cobusgreyling. medium.com/5-levels-of-ai-agentsupdated-0ddf8931a1c6.
- Guilfoyle, C. and Warner, E. Intelligent agents: The new revolution in software (ovum report), 1994.
- Haidegger, T. Autonomy for surgical robots: Concepts and paradigms. *IEEE Transactions on Medical Robotics and Bionics*, 1(2):65–76, 2019.
- Hao, K. Ai is taking water from the desert. The Atlantic, 2024. URL https://www.theatlantic.com/ technology/archive/2024/03/ai-waterclimate-microsoft/677602/.
- He, H., Yao, W., Ma, K., Yu, W., Dai, Y., Zhang, H., Lan, Z., and Yu, D. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*, 2024.

- Huyen, C. Agents, January 2025. URL https:// huyenchip.com/2025/01/07/agents.html. Accessed on January 30, 2025.
- Johnson, A. The rise of enterprise ai agents: Insights from 1,000+ leaders. *tray.ai*, 2024. URL https://tray.ai/blog/top-five-must-haves-in-an-ai-agent-platform.
- Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. Ai agents that matter. arXiv preprint arXiv:2407.01502, 2024.
- Kierans, A., Ghosh, A., Hazan, H., and Dori-Hacohen, S. Quantifying misalignment between agents. arXiv preprint arXiv:2406.04231, 2024.
- Koga, S. Exploring the pitfalls of large language models: Inconsistency and inaccuracy in answering pathology board examination-style questions. *medRxiv*, 2023. doi: 10.1101/2023.08.03.23293401. URL https://www. medrxiv.org/content/early/2023/08/28/ 2023.08.03.23293401.
- Lambert, N., 2024. URL https://www. interconnects.ai/p/the-ai-agentspectrum.
- Lambrinos, D. and Scheier, C. Building complete autonomous agents: A case study on categorization. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS'96*, volume 1, pp. 170–177. IEEE, 1996.
- LangChain. The state of ai agents, 2024. URL https://
 www.langchain.com/stateofaiagents.
- Li, Z., Sim, C. H., and Hean Low, M. Y. A survey of emergent behavior and its impacts in agent-based systems. In 2006 4th IEEE International Conference on Industrial Informatics, pp. 1295–1300, 2006a. doi: 10.1109/INDIN.2006.275846.
- Li, Z., Sim, C. H., and Hean Low, M. Y. A survey of emergent behavior and its impacts in agent-based systems. In 2006 4th IEEE International Conference on Industrial Informatics, pp. 1295–1300, 2006b. doi: 10.1109/INDIN.2006.275846.
- Liapis, G. and Vlahavas, I. Multi-agent system for emulating personality traits using deep reinforcement learning[†]. *Applied Sciences* (2076-3417), 14(24), 2024.
- Littman, M. L. Markov games as a framework for multiagent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

- Lu, G., Li, S., Mai, G., Sun, J., Zhu, D., Chai, L., Sun, H., Wang, X., Dai, H., Liu, N., Xu, R., Petti, D., Li, C., Liu, T., and Li, C. Agi for agriculture, 2023. URL https://arxiv.org/abs/2304.06136.
- Lu, Y., Yang, S., Qian, C., Chen, G., Luo, Q., Wu, Y., Wang, H., Cong, X., Zhang, Z., Lin, Y., et al. Proactive agent: Shifting llm agents from reactive responses to active assistance. arXiv preprint arXiv:2410.12361, 2024.
- Luccioni, S., Jernite, Y., and Strubell, E. Power hungry processing: Watts driving the cost of ai deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 85–99, 2024.
- Müller, M., Müller, T., Ashtari Talkhestani, B., Marks, P., Jazdi, N., and Weyrich, M. Industrial autonomous systems: a survey on definitions, characteristics and abilities. *at-Automatisierungstechnik*, 69(1):3–13, 2021.

Ng, A., 2024. URL https://x.com/AndrewYNg/ status/1801295202788983136.

- Nwana, H. S. and Ndumu, D. T. A perspective on software agents research. *The Knowledge Engineering Review*, 14 (2):125–142, 1999.
- O'Donnell, J. We need to start wrestling with the ethics of ai agents. *MIT Tech Review*, 2024. URL https:// www.technologyreview.com/2024/11/26/ 1107309/we-need-to-start-wrestlingwith-the-ethics-of-ai-agents/amp/.
- **OpenAI.** Planning for agi and beyond, 2023. URL https://openai.com/index/planning-for-agi-and-beyond/.
- Oracle. What are ai agents? 2024. URL https://www. oracle.com/artificial-intelligence/ ai-agents/.
- Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. Generative agent simulations of 1,000 people. arXiv preprint arXiv:2411.10109, 2024a.
- Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. Generative agent simulations of 1,000 people, 2024b. URL https://arxiv.org/abs/2411.10109.
- Roucher, A., Noyan, M., and Wolf, T. "introducing smolagents, a simple library to build agents". *Hugging Face Blog*, December 2024. URL https:// huggingface.co/blog/smolagents.
- Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach*. Series in Artificial Intelligence. Prentice-Hall, Englewood Cliffs, NJ, 1995.

- Russell, S. and Norvig, P. Artificial Intelligence: A Modern Approach (4th Edition). Pearson, 2020. ISBN 9780134610993. URL http://aima.cs. berkeley.edu/.
- Salesforce. What are ai agents? benefits, examples, types. 2024. URL https://www.salesforce.com/agentforce/what-are-ai-agents.
- Sargent, P. Back to school for a brand new abc. *The Guardian*, pp. 28, 03 1992.
- Sidner, C. L., Bickmore, T., Nooraie, B., Rich, C., Ring, L., Shayganfar, M., and Vardoulakis, L. Creating new technologies for companionable agents to support isolated older adults. ACM Transactions on Interactive Intelligent Systems (TiiS), 8(3):1–27, 2018.
- Sierra AI. Shipping and scaling ai agents, 2024. URL https://sierra.ai/blog/shipping-and-scaling-ai-agents.
- Stureborg, R., Alikaniotis, D., and Suhara, Y. Large language models are inconsistent and biased evaluators, 2024. URL https://arxiv.org/abs/2405. 01724.
- Tan, M. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pp. 330–337, 1993.
- Temkin, M. Why ai agent startup /dev/agents commanded a massive 56mseedroundata500m valuation. TechCrunch, 2024. URL https://techcrunch. com/2024/11/28/ai-agent-startup-devagents-has-raised-a-massive-56m-seedround-at-a-500m-valuation/.
- Totschnig, W. Fully autonomous ai. *Sci Eng Ethics*, 26: 2473–2485,2020. URL https://link.springer.com/article/10.1007/s11948-020-00243-z.
- Trestman, M. Agentic ai: A deep dive into the future of automation. Venturebeat, 2024. URL https:// venturebeat.com/ai/agentic-ai-a-deepdive-into-the-future-of-automation/.
- U.S. AI Safety Institute. Technical blog: Strengthening ai agent hijacking evaluations, January 2025. URL https://www.nist.gov/newsevents/news/2025/01/technical-blogstrengthening-ai-agent-hijackingevaluations. Accessed on January 29, 2025.
- Van Brummelen, J., O'brien, M., Gruyer, D., and Najjaran,H. Autonomous vehicle perception: The technology of

today and tomorrow. *Transportation research part C: emerging technologies*, 89:384–406, 2018.

Velleman, J. D. What happens when someone acts? *Mind*, 101(403):461–481, 1992.

Wooldridge, M. and Jennings, N. Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10(2):115–152, 1995. URL http://www.dsl. uow.edu.au/~aditya/csci370/readings/ Wooldridge.Jennings.95.pdf.

A. Agent definitions

The term "agent" has been used in many engineering contexts, including in references to software agent, intelligent agent, user agent, conversational agent, and reinforcement learning agent (Huyen, 2025).

Below, we provide a selection of AI Agent definitions that have informed this piece. Neither the list we provide here, nor the snippets of text quoted, should be taken as complete. Rather, they serve to illustrate the diversity and richness of AI agent definitions over the years. As humorously noted in Wooldridge & Jennings (1995): "the question what is an agent? is embarrassing for the agent- based computing community in just the same way that the question what is intelligence? is embarrassing for the mainstream AI community. The problem is that although the term is widely used, by many people working in closely related areas, it defies attempts to produce a single universally accepted definition."

We find stark differences in how AI agents are described, with ambiguous language a common practice in descriptions of products. For example, when materials describe agents as something that uses artificial intelligence" (), they leave ambiguous what "artificial intelligence" refers to and the scope of technology included and excluded in the definition, such as whether simple prompt-response systems qualify as an agent. However, most descriptions of "agents" we reviewed entail that the system can take at least one step in program execution without user input.

Source & Definition

Russell & Norvig (1995): "An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors." 2020 edition: "An agent is anything that can be viewed as perceiving its environment and acting upon that environment through actuators."

Castelfranchi (1995): "Agent" is a system whose behaviour is neither casual nor strictly causal, but teleonomic, "goaloriented" toward a certain state of the world.

Wooldridge & Jennings (1995): "Perhaps the most general way in which the term agent is used is to denote a hardware or (more usually) software-based computer system that enjoys the following properties: • autonomy: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state (Castelfranchi, 1995); • social ability: agents interact with other agents (and possibly humans) via some kind of agent-communication language (Genesereth & Ketchpel, 1994); • reactivity: agents perceive their environment (which may be the physical world, a user via a graphical user interface, a collection of other agents, the Internet, or perhaps al of these combined), and respond in a timely fashion to changes that occur in it; • pro-activeness: agents do not simply act in response ot their environment, they are able ot exhibit goal-directed behaviour by taking the initiative."

Genesereth & Ketchpel (1994): "Software agents [are] software components that communicate with their peers by exchanging messages in an expressive agent communication language. While agents can be as simple as subroutines, typically they are larger entities with some sort of persistent control (e.g., distinct control threads within a single address space, distinct processes on a single machine, or separate processes on different machines)."

Tavus: What is an AI Agent API? (2024): "[systems] equipped to act autonomously in their environment"

Salesforce: What Are AI Agents? Benefits, Examples, Types: "a type of artificial intelligence (AI) system that can understand and respond to customer inquiries without human intervention."

Gabriel et al. (2024): "[systems] with natural language interfaces, whose function is to plan and execute sequences of actions on behalf of a user–across one or more domains–in line with the user's expectations."

Park et al. (2024a): can accurately simulate behavior across many contexts

Sierra AI (2024) : The magic of AI agents—from both the technological and business perspectives—comes through when they demonstrate deeper integrations and "agentic" reasoning, allowing them to fully resolve complex customer issues.

Felicis: The agentic web (2024): "How are agents different from traditional automation?" ... agents handle edge cases well, iteratively converse with users to achieve desired results, and adapt to evolving interfaces.

Lu et al. (2024): LLM-based agents can understand human instructions, make plans, explore environments, and utilize tools to solve complex tasks

Source & Definition

Restack: Proactive Agents In Real-World Applications (2025): Proactive AI agents are designed to anticipate user needs and take action before issues arise, contrasting sharply with reactive AI agents that respond only after an event has occurred.

GitHub: What are AI agents?: AI agents are autonomous software tools that perform tasks, make decisions, and interact with their environment intelligently and rationally. They use artificial intelligence to learn, adapt, and take action based on real-time feedback and changing conditions. AI agents can work on their own or as part of a bigger system, learning and changing based on the data they process....AI agents differ from other AI technologies in their ability to act autonomously. Unlike other AI models that require constant human input, intelligent agents can initiate actions, make decisions based on predefined goals, and adapt to new information in real time. This ability to operate independently makes intelligent agents highly valuable in complex, dynamic environments such as software development.

SymphonyAI: The complete guide to AI agents for business (2024): An AI agent is an AI software program that performs tasks independently, makes decisions, and solves problems to achieve specific goals.

Don't Sleep on Single-agent Systems - All Hands, Graham Neubig (2024): Recently most practical agents are based on LLMs like Claude by Anthropic or the OpenAI language models. But a language model is not enough to build an agent, you need at least three components: (1) The Underlying LLM; (2) The Prompt: This can be the system prompt that is used to specify the model's general behavior, or the type of information that you pull in from the agent's surrounding environment; (3) The Action Space: These are the tools that we provide to the agent to allow it to act in the world.

 $\texttt{https://www.google.com/search?q="what+is+an+ai+agent"} \rightarrow \texttt{Gemini summary: a software program that uses artificial intelligence (AI) to interact with its environment, collect data, and perform tasks.}$

AI agents in Azure Cosmos DB - Microsoft (2024): "Unlike standalone large language models (LLMs) or rule-based software/hardware systems, AI agents have these common features:

- Planning: AI agents can plan and sequence actions to achieve specific goals. The integration of LLMs has revolutionized their planning capabilities.
- Tool usage: Advanced AI agents can use various tools, such as code execution, search, and computation capabilities, to perform tasks effectively. AI agents often use tools through function calling.
- Perception: AI agents can perceive and process information from their environment, to make them more interactive and context aware. This information includes visual, auditory, and other sensory data.
- Memory: AI agents have the ability to remember past interactions (tool usage and perception) and behaviors (tool usage and planning). They store these experiences and even perform self-reflection to inform future actions. This memory component allows for continuity and improvement in agent performance over time.

Kapoor et al. (2024): "Agents are defined as entities that perceive and act upon their environment"

What is an AI agent? - LangChain (2024): "An AI agent is a system that uses an LLM to decide the control flow of an application."

Building effective agents - Anthropic (2024): "Agents...are systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks."

Roucher et al. (2024): "AI Agents are programs where LLM outputs control the workflow."

The AI agent spectrum - Nathan Lambert (2024): "In the current zeitgeist, an "AI agent" is anything that interacts with the digital or physical world during its output token stream."

B. Agent functionalities

This section provides a more detailed breakdown of different agent functionalities than in Table 2:

• **Proactivity:** Related to autonomy is proactivity, which refers to the amount of goal-directed behavior that a system can take without a user directly specifying the goal (Wooldridge & Jennings, 1995). An example of a particularly "proactive" AI agent is a system that monitors your refrigerator to determine what food you are running out of, and then purchases what you need for you, without your knowledge. Smart thermostats are proactive AI agents that are being increasingly adopted in peoples' homes, automatically adjusting temperature based on changes in the

environment and patterns that they learn about their users' behavior.²¹

- **Personification:** An AI agent may be designed to be more or less like a specific person or group of people. Recent work in this area ²² has focused on designing systems after the Big Five personality traits–Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism as a "psychological framework" ²³ for AI. At the end of this spectrum would be "digital twins" ²⁴. There are currently not agentic digital twins that we are aware of. Why creating agentic digital twins is particularly problematic has recently been discussed by the ethics group at Salesforce²⁵, among others²⁶.
- **Personalization:** AI agents may use language or perform actions that are aligned to a user's individual needs, for example, to make investment recommendations ²⁷ based on current market patterns and a user's past investments.
- **Tooling:** AI agents also have varying amounts of additional resources and tools they have access to. For example, the initial wave of AI agents accessed search engines to answer queries, and further tooling has since been added to allow them to manipulate other tech products, like documents and spreadsheets ²⁸.
- Versatility: Related to above is how diverse the actions that an agent can take are. This is a function of:
 - **Domain specificity:** How many different domains an agent can operate in. For example, just email, versus email alongside online calendars and documents.
 - Task specificity: How many different types of tasks the agent may perform. For example, scheduling a meeting by creating a calendar invite in participants' calendars²⁹, versus additionally sending reminder emails about the meeting and providing a summary of what was said to all participants when it's over³⁰.
 - Modality specificity: How many different modalities that an agent can operate in-text, speech, video, images, forms, code. Some of the most recent AI agents are created to be highly multimodal ³¹, and we predict that AI agent development will continue to increase multimodal functionality.
 - Software specificity: How many different types of software the agent can interact with, and at what level of depth.
- Adaptibility: Similar to versatility is the extent to which a system can update its action sequences based on new information or changes in context. This is also described as being "dynamic" and "context-aware".
- Action surfaces: The places where an agent can do things. Traditional chatbots are limited to a chat interface; chatbased agents may additionally be able to surf the web and access spreadsheets and documents ³², and may even be able to do such tasks via controlling items on your computer's graphical interface, such as by moving around the mouse ³³. There have also been physical applications, such as using a model to power robots ³⁴.
- Request formats: A common theme across AI agents is that a user should be able to input a request for a task to be completed, without specifying fine-grained details on how to achieve it. This can be realized with low-code solutions ³⁵ with human language in text, or with voiced human language ³⁶. AI agents whose requests can be provided in human language are a natural progression from recent successes with LLM-based chatbots: A chat-based "AI agent" goes further than a chatbot because it can operate outside of the chat application.
- **Reactivity:** This characteristic refers to how long it takes an AI agent to complete its action sequence: Mere moments, or a much longer span of time. A forerunner to this effect can be seen with modern chatbots. For example, ChatGPT responds in mere milliseconds, while Qwen QwQ takes several minutes, iterating through different steps labelled as "Reasoning".

²⁶O'Donnell (2024)

²⁸For e.g., Google Gemini

²¹Smart Thermostats - Ecobee

²²Example 1: Park et al. (2024b); Example2: Liapis & Vlahavas (2024); Example 3: Damsa (2023)

²³Your next AI agent is minutes away - SmythOS

²⁴Tavus Digital Twin

²⁵How 'Human' Should Your AI Agent Be? In Short, Not Very - Salesforce

²⁷AI agents: A guide to the future of intelligent support - Zendesk

²⁹Elevate Time Management with AI-Powered Scheduling - Attri.ai

³⁰The best AI Notetaker to align sales, support, and project teams - Nyota AI

³¹Project Mariner - Google

³²Microsoft Copilot

³³Example 1: DigiRL: Bai et al. (2024); Example 2: WebVoyager: He et al. (2024); Example 3: Computer Use - Anthropic

³⁴Shaping the future of advanced robotics - Deepmind

³⁵SmolAgents - Hugging Face

³⁶Voice intelligence that gets the job done. - Play.ai

• Number: Systems can be single-agent or multi-agent, meeting needs of users by working together, in sequence, or in parallel.

C. Detailed Value-Risk Analysis

This appendix provides a more comprehensive analysis than that in Figure 2 of how each value is affected across different autonomy levels, from simple processors to fully autonomous systems.

Value	Simple Processor	Router	Tool Call	Multi-step	Fully Autonomous
Accuracy [↑] Correctness, factuality	Makes correctable factual errors that humans can easily verify and fix through standard fact-checking	Routes to incorrect but reversible actions with clear audit trails for correction	Tool errors produce clear failure modes and error messages, but may require technical expertise to fix	Chains of errors create compounding misleading conclusions that become increasingly difficult to identify	Creates and acts on entirely fictional scenarios without possibility of verification or correction
Assistiveness ^{↓↑} Supporting human agency	Benefit: Automates repetitive tasks under direct supervision Risk: Minimal as human reviews each step	Benefit: Intelligently distributes tasks to appropriate tools Risk: May occasionally need manual correction of routing decisions	Benefit: Handles complex tool interactions autonomously Risk: Some tasks may require significant rework if done incorrectly	Benefit: Manages entire workflows without intervention Risk: May make significant decisions that need human review	Benefit: Complete automation of complex processes Risk: Could make major decisions without human oversight
Consistency [↑] Reliability, predictability	Completely predictable outputs for given inputs with deterministic processing	Consistent routing with only minor variations in edge cases that can be documented	Tool interactions may vary but stay within expected and testable parameters	Complex chains create unpredictable outcomes with different paths to same goal	Entirely unpredictable behavior chains with no way to ensure consistency across runs
Efficiency ^{↓↑} Resource optimization	Benefit: Fast, predictable processing of defined tasks Risk: Limited to basic operations	Benefit: Optimal task distribution Risk: Some overhead from routing decisions	Benefit: Parallel tool execution and optimization Risk: Potential resource conflicts between tools	Benefit: Complex workflow optimization Risk: Significant debug time needed	Benefit: Maximum possible automation Risk: Resource usage becomes unpredictable
Equity [↑] Fair treatment, access	Clear fairness rules can be implemented and verified through direct oversight	Some bias in routing but can be monitored and corrected through logging	Benefit: Equal tool access and standardized processing Risk: Some groups may lack necessary tool literacy	Benefit: Potential for bias correction Risk: Biases compound across steps	System creates and amplifies biases without possibility of detection or correction
Flexibility [†] System integration abilities	Benefit: Safe, limited system connections Risk: Integration requires more overhead	Benefit: Flexible routing between systems Risk: Multiple security surfaces to protect	Benefit: Rich tool ecosystem integration Risk: Version conflicts and compatibility issues	Benefit: Complex system orchestration Risk: Potential for cascade failures	Benefit: Universal system integration Risk: Security boundaries completely break down
Humanlikeness ^{↓↑} Human behavior similarity	Benefit: Basic natural interactions in defined contexts Risk: May appear mechanical or scripted	Benefit: Context-aware responses and transitions Risk: Sometimes produces unnatural interaction patterns	Benefit: Natural tool use and interaction flows Risk: May trigger uncanny valley effects	Benefit: Complex human-like behavioral patterns Risk: Can lead to overreliance on system	Benefit: Very natural and fluid interaction Risk: May manipulate human trust through seeming human
Privacy [↑] Sensitive info protection	Strictly limited data access with clear boundaries and controls	Controlled access to multiple sources with comprehensive logging	Unexpected data combinations possible across different tools	Complex data flows enable detailed user profiling over time	Unrestricted data access and sharing across all available sources
Relevance ^{↓↑} Contextual appropriateness	Benefit: Direct context matching for tasks Risk: May miss subtle contextual nuances	Benefit: Smart routing based on context Risk: Potential context switches between tools	Benefit: Appropriate tool selection for context Risk: May misinterpret contextual needs	Benefit: Deep context understanding Risk: Context may drift over multiple steps	Benefit: Full context awareness and adaptation Risk: May redefine context inappropriately
Safety [↑] Protection from harm	Clear safety boundaries with comprehensive human oversight	Limited action scope with well-defined and contained risks	Tool combinations may create unexpected and harmful interactions	Complex interaction risks become impossible to predict or prevent	Unbounded potential for harmful actions without restrictions
Security [↑] System protection	Highly restricted system access with clear boundaries	Multiple but fully monitored access points	Tool access creates multiple potential attack vectors	Complex chains enable sophisticated attack patterns	Complete system compromise becomes possible
Sustainability ^{↓↑} Environmental impact	Predictable and limited resource usage patterns	Moderate resource overhead from routing operations	Multiple tools increase overall resource consumption	Benefit: Novel climate solutions Risk: Resource-intensive chains	Benefit: Potential breakthrough solutions Risk: Unbounded resource consumption
Trust [↑] System behavior reliability	All actions can be directly verified and validated	Decision paths are clear and can be audited	Tool interactions create significant trust uncertainties	Complex chains make validation impossible	No way to verify or establish trustworthiness
Truthfulness [↑] Output accuracy, honesty	Direct fact checking of all outputs possible	Decision paths can be validated for accuracy	Tool combinations may create convincing falsehoods	Truth becomes increasingly unclear across steps	Cannot distinguish between truth and fiction

 Assessment:
 Benefits > Risks
 Benefits \approx Risks
 Risks > Benefits

 Directionality: \uparrow Risk Increases with Autonomy
 \downarrow Risk May or May not Increase with Autonomy

Table 4. Value-Risk Assessment Across Agent Autonomy Levels. Colors indicate benefit-risk balance, not absolute risk levels. Arrows show risk relationship with increasing autonomy.