

Supplement on AI Agents, Cybersecurity & the Human at the Keyboard

Preparing Students, Faculty & Staff to Use AI Wisely

• A Framework for the AI-Ready University •



AI LITERACY
Use with Understanding



CYBERSECURITY
Protect What Matters



ETHICS
Act with Integrity



EDUCATION
Form Human Flourishing

*“Technology should expand human capability
without diminishing human responsibility.”*



Guidebook Supplement:

Who's Responsible for AI Safety on Your Campus?

That raises a bigger question:

Who actually owns AI safety on your campus?

IT? Academic Affairs? Information Security? The Cabinet?

Increasingly, the answer is all of them.

But there is another place where AI safety begins, and it may be the most important one: **The first line of security is fingers on the keyboard.**

Every AI interaction begins with a human decision: what to enter, what to trust, what to question, and what to share.

Institutional safeguards matter. Policies matter. Technology matters. But none of them replaces the judgment of the person sitting at the keyboard.

AI safety is therefore not simply an IT issue, a cybersecurity issue, or an academic policy issue. It is an institutional responsibility that ultimately depends on human judgment and responsible behavior.

AI literacy is ultimately a human formation issue, not merely a technology-training issue.

Why This Matters:

AI isn't waiting for your institution to catch up.

Students, faculty, and staff are already using AI, often faster than institutions can establish the policies, safeguards, and accountability needed to manage it responsibly.

The next stage of artificial intelligence is not simply about AI that can answer questions. Increasingly, AI systems are being developed that can act.

An AI agent may be able to search, retrieve information, interact with software, make decisions within defined parameters, and carry out tasks on behalf of a user. That creates enormous possibilities for productivity and learning. It also introduces a new cybersecurity challenge: **What happens when AI moves from giving an answer to taking an action?**

The distinction is important.

A traditional chatbot generally waits for a person to ask something and then produces a response. **An AI agent can potentially take a series of actions with less direct human involvement.**

That changes the cybersecurity equation.

The university therefore needs to think not only about protecting people from AI, but also about protecting people, systems, data, and institutions when AI has the ability to act.



PART I From AI That Answers to AI That Acts.

1. The Changing Nature of AI

The development of increasingly agentic AI represents a significant shift.

AI is moving along a spectrum:

AI that responds → AI that assists → AI that acts

Each step introduces a different level of responsibility and risk.

An AI system that produces an incorrect answer can cause a person to make a bad decision. An AI system that can independently take action introduces another possibility: **The system can make the bad decision and act on it.**

This does not mean that AI agents are inherently dangerous. It means that the design of the surrounding environment becomes increasingly important.

The central cybersecurity question becomes:

What authority have we given the AI, and what happens if that authority is misused, misunderstood, or compromised?



PART II. The New AI Cybersecurity Problem

1. More Capability Means More Responsibility

The more an AI system can do, the more carefully its permissions need to be controlled. An agent should not automatically receive broad access simply because it is capable of using that access.

This leads to a fundamental cybersecurity principle:

Capability does not require unlimited authority.

An AI agent may be capable of reading files, accessing systems, communicating with external services, or executing tasks. **That does not mean it should be permitted to do all of those things.**

The university must determine:

- What the agent is allowed to access
- What information it is allowed to see
- What systems it is allowed to interact with

- What actions require human approval
- How long its permissions remain valid
- How its activity is monitored
- How its access can be revoked



PART III. Containing the Agent

1. Sandboxing

One important protection is sandboxing.

An AI agent can be placed in a controlled environment where its ability to interact with the larger institutional environment is limited.

The principle is simple:

Give the agent room to work, but not unlimited room to roam.

If an agent makes an error, becomes compromised, or behaves unexpectedly, containment can prevent that error from becoming a much larger institutional problem.

2. Least Privilege

The principle of least privilege becomes especially important with agentic AI. An agent should receive only the permissions necessary to perform its assigned task.

If an agent needs access to one database, it should not automatically receive access to every database. If it needs to perform one task, it should not automatically receive authority to perform unrelated tasks.

The question should always be:

What is the minimum authority this AI needs to accomplish its job?

3. Restricted Network Access

AI agents should not automatically have unrestricted access to institutional networks or external services. Restricting network access can limit the consequences of an error or compromise.

The goal is not to prevent useful activity.

The goal is to establish boundaries.



IV. Monitoring, Logging & Revocation

1. Trust, But Verify

An AI agent should not become a technological employee whose activities nobody watches. **Independent monitoring can provide visibility into what an agent is doing.**

Logging can help answer questions such as:

- What did the agent access?
- What actions did it take?
- What information did it transmit?
- What decisions did it make?
- When did something unusual begin happening?

Monitoring becomes particularly important when an agent has access to consequential systems.

2. Short-Lived Credentials

Long-lasting credentials create unnecessary risk.

Where possible, access should be limited in both scope and duration.

An agent may need permission to perform a particular task for a particular period of time. Once the task is complete, that access should expire.

3. The Ability to Pull the Plug

Every agentic system needs a way to have its authority rapidly revoked.

If an agent begins behaving unexpectedly, the institution should not have to negotiate with the agent. It needs a way to stop it.

That means having mechanisms for:

- Revoking credentials
- Ending sessions
- Isolating systems
- Blocking network access
- Suspending the agent
- Terminating the process when necessary

The ability to stop an AI system is not a sign that the system has failed. It is part of responsible system design.



PART V. The Human in the Loop

1. Human Approval for Consequential Actions

Not every AI action requires a person to approve every click. But consequential actions may require meaningful human oversight.

There is a difference between an AI agent: drafting an email and sending the email.

There is a difference between: identifying a possible security problem and changing an institutional security configuration.

There is a difference between: recommending an action and taking the action.

The more consequential the action, the more important meaningful human oversight becomes.

2. The Fingers on the Keyboard

This returns us to one of the central ideas of this Guidebook:

Who owns the AI cybersecurity on your campus?

The answer cannot simply be IT.

It cannot simply be the CISO.

It cannot simply be faculty, administrators, or the vendor supplying the technology.

Ultimately, **cybersecurity involves people making decisions.**

The first line of security is often: The fingers on the keyboard.

Someone decides what information to enter.

Someone decides what to authorize.

Someone decides what to click.

Someone decides what permissions to grant.

Someone decides whether an AI-generated request is legitimate.

Someone decides whether an AI recommendation should be trusted.

AI can increase the speed and scale of what happens next.

But human beings still have responsibility for the decisions that put those systems into motion.



PART VI. The AI “Genie” Problem

1. Can We Put the Genie Back in the Bottle?

As AI becomes more capable and more widely distributed, another question emerges:

What happens if powerful AI capability becomes widely available before society has developed adequate safeguards?

Once a capability is distributed across many systems, organizations, and individuals, it **becomes increasingly difficult to imagine simply putting it back in the bottle.**

The challenge therefore cannot be reduced to: “How do we make every AI perfectly safe?”

A more realistic question is: How do we build systems and institutions that remain resilient even when some AI systems behave badly?

That changes the objective.

The goal becomes not merely prevention,

but also: containment + detection + response + recovery

2. When the Genie Must Be Contained

An unsafe or compromised AI system may need to be isolated or terminated. That is not necessarily an admission that AI has become uncontrollable.

It is an application of a familiar cybersecurity principle:

When a system becomes dangerous, contain it and, when necessary, shut it down.

The important question is whether the institution has the authority and technical capability to do so quickly.



PART VII. From Perfect Safety to Resilience

1. The New Security Mindset

No cybersecurity strategy can guarantee that every attack, error, or misuse will be prevented. The objective is to make sure that when something goes wrong, the consequences remain manageable.

For agentic AI, that means designing systems so that: A mistake does not become a catastrophe.

- A compromised agent does not automatically compromise the entire institution.
- A stolen credential does not provide unlimited access.
- An erroneous decision does not automatically become an irreversible action.
- An AI system that behaves unexpectedly can be isolated.
- An unsafe process can be stopped.
- The institution can recover.

2. Resilience as AI Literacy

This is where cybersecurity and AI literacy meet. Technical safeguards matter enormously.

But people still need to understand:

- **What AI is capable of doing**
- **What authority an AI system has been given**
- **What information it can access**
- **What risks its actions create**
- **When human approval is necessary**
- **How to recognize unusual behavior**

- **When to stop an AI-assisted process**
- **Who is responsible for responding**

AI literacy therefore includes more than knowing how to use AI.

It includes knowing when not to trust it, when to question it, and when to stop it.



PART VIII. The Human Responsibility Remains

The emergence of agentic AI does not eliminate human responsibility. It makes human responsibility more important.

The fundamental question is not: “Can AI do this?”

The more important questions are:

“Should AI do this?”

“How much authority should AI have?”

“Who remains responsible when AI acts?”

And perhaps most importantly:

“Have we designed the system so that a human being can still exercise meaningful judgment?”

That is the deeper purpose of AI literacy.

We are not simply teaching people how to operate increasingly powerful technology.

We are preparing them to remain responsible when increasingly powerful technology operates around them.



PART IX. Practical Framework for the AI-ready University

As universities encounter increasingly agentic AI, five questions can provide a practical starting point:

1. AUTHORITY

What is this AI allowed to do?

2. ACCESS

What information and systems can it reach?

3. AUTONOMY

What decisions can the AI make on its own?

4. OVERSIGHT

Which actions require human judgment or approval?

5. CONTAINMENT

What prevents an error or compromise from spreading?

6. SHUTDOWN

How quickly can we stop the system if something goes wrong?

These questions should become part of institutional AI governance, cybersecurity planning, procurement, faculty guidance, and AI literacy education.



PART X. The Larger Lesson

AI may become increasingly autonomous.

But autonomy does not have to mean abandonment of human responsibility.

The goal is not to create AI systems that require no human judgment.

The goal is to create systems in which human judgment remains meaningful even when AI is highly capable.

The university's responsibility is therefore twofold:

Build secure systems.

Form responsible people.

Technical safeguards protect the institution from technological failure.

Human formation helps protect the institution from human failure.

We need both.

Because ultimately, the most sophisticated cybersecurity system on campus still depends upon a human being somewhere in the chain of action.

The fingers are still on the keyboard.

And that is where AI literacy, cybersecurity, judgment, and responsibility finally meet.



PART XI. ChatGPT vs. Agentic AI

From "What Can You Tell Me?" to "What Can You Do for Me?"

One of the most important developments in artificial intelligence is the movement from AI that primarily responds to AI that can increasingly act.

Understanding the difference is essential for understanding the emerging cybersecurity challenges surrounding AI.

1. In its simplest form, ChatGPT is a conversational AI.

A person asks a question, makes a request, or provides a task.

For example: "Explain the difference between phishing and spear-phishing."
ChatGPT generates a response.

Or: "Help me write an email."

ChatGPT generates a draft.

The basic pattern is:

You → AI → Response → You decide what happens next

"In a traditional conversational interaction, the human remains the primary driver of the interaction."

ChatGPT can help a person think, research, write, analyze, brainstorm, explain, and solve problems. But in the traditional conversational model, the person remains responsible for deciding what to do with the response.

2. Agentic AI: "Go Do Something"

Agentic AI introduces another level of capability.

Instead of simply responding to a request, **an AI agent may be given an objective and the ability to take multiple steps toward accomplishing it.**

For example:

"Monitor our university's security alerts, investigate anything suspicious, determine whether it appears to be a phishing attempt, check the relevant systems, and notify the security team if the evidence meets our criteria."

An agent could potentially:

- 1. Receive the objective.**
- 2. Break the objective into tasks.**
- 3. Search systems and retrieve information.**
- 4. Make intermediate decisions.**
- 5. Use software, websites, or APIs.**
- 6. Take actions.**
- 7. Evaluate the results.**
- 8. Continue until the task is completed or it reaches a defined boundary.**

The pattern becomes:

You → AI Agent → Decisions → Tools → Actions → Results

That is a significant change.

3. The Critical Difference: Agency

The distinction between conversational AI and agentic AI is not simply intelligence. It is agency.

A conversational AI primarily helps you determine what to do.

An agentic system may be capable of helping determine how to do it and then doing it.

The more autonomous an AI system has, the more important questions of authority, access, oversight, and accountability become.

This leads to a fundamental cybersecurity question: What authority have we given the AI, and what happens if that authority is misused, misunderstood, or compromised?

4. The Keyboard Analogy

This distinction connects directly to one of the central ideas of this Guidebook:
Who owns the AI cybersecurity on your campus?

Our answer is: The fingers on the keyboard.

With ordinary conversational AI, those fingers are usually quite literal.

You ask:

AI responds.

You evaluate.

You decide.

You click.

With an agentic AI agent, however, the situation changes.

You may give the agent an assignment and the authority to carry out much of the work.

The agentic AI may have access to:

- Files
- Websites
- Applications
- Databases
- APIs
- Email
- Institutional systems
- Credentials

The human may still have established the original objective, but **the AI is now participating directly in the chain of action.**

That raises a new question: Whose authority are those fingers exercising?

5. A Simple Analogy

Think of ChatGPT as a knowledgeable consultant sitting beside you.

You ask: “What should I do?”

The consultant gives you advice.

You decide whether to follow it.

An Agentic AI agent is closer to saying:

“Here’s the assignment. Here’s your badge. Here’s access to the office. Go take care of it.”

Now different questions become necessary:
What doors does the badge open?
What is the agent authorized to do?
Who is watching?
What happens if it makes a mistake?
Can we take the badge away?

Those are cybersecurity questions, but they are also questions of human responsibility.

6. Why Agentic AI Changes the Cybersecurity Equation

An incorrect answer from a **conversational AI** can certainly cause harm. A person might believe the answer, act upon it, and make a poor decision.

But **an agentic system** can potentially move from error to action much more quickly.

The distinction can be summarized this way:

Conversational AI	Agentic AI
Answers questions	Pursues objectives
Generates responses	Takes actions
Human commonly directs the interaction	AI may determine intermediate steps
Human commonly perform the next action	AI may perform the next action
Primarily informational	Potentially operational
Risk centers on what the AI says	Risk includes what the AI does

The dividing line is not absolute. Modern AI systems can combine conversational and agentic capabilities.

The important issue is the degree of autonomy, tool access, decision-making, and authority given to the system.

7. The Question Universities Need to Ask

These questions should become part of institutional AI governance, cybersecurity planning, procurement, faculty guidance, and AI literacy education.

**AI moves from: "What can AI tell me?"
to: "What can AI do for me?"**

And eventually: “What can AI do without asking me first?”

That progression should **immediately trigger questions about:**

1. AUTHORITY

What is the AI allowed to do?

2. ACCESS

What information and systems can it reach?

3. AUTONOMY

Which decisions can it make on its own?

4. OVERSIGHT

Which actions require meaningful human approval?

5. CONTAINMENT

What prevents an error or compromise from spreading?

6. SHUTDOWN

How quickly can we stop it if something goes wrong?

8. ChatGPT Is Not the Enemy

The distinction between ChatGPT and agentic AI should not be presented as safe AI versus dangerous AI. That is too simplistic.

Both conversational and agentic systems can be useful. Both can create risks. The important question is not simply what the technology is called.

The important questions are:

- What can it access?
- What can it decide?
- What can it do?
- What authority have we given it?
- What safeguards surround it?

This is why AI literacy cannot stop at learning how to use an AI application.

Users need to understand the capabilities and boundaries of the particular AI system they are using.

9. The Bottom Line

The movement toward agentic AI represents a significant change in the relationship between humans and artificial intelligence.

With conversational AI, we commonly ask: “What can you tell me?”

With agentic AI, we increasingly ask: “What can you do for me?”

The second question carries greater responsibility.

- **As AI gains the ability to act, institutions must become increasingly deliberate about authority, access, oversight, containment, and accountability.**

The goal is not to prevent AI from becoming useful.

The goal is to ensure that greater AI capability does not mean less human responsibility.

And that brings us back to the central principle of AI literacy:

AI may have produced the action, but human beings and institutions remain responsible for the authority they give it.

The fingers are still on the keyboard.

That is where AI literacy, cybersecurity, judgment, and responsibility finally meet.