Class Meeting 2

Professional Judgment, Stakeholders & AI-Enabled Cybersecurity

Cyber Ethics & Law

Today's Agenda

Section 1

Professional Codes of Ethics
Competing Professional Responsibilities
Stakeholder Analysis

Section 2

Privacy, Authorization & Surveillance
Case Introduction: AI Flags an Insider Threat

Section 3

AI-Assisted Threat Detection
False Positives, Bias & Assumptions
Human Oversight & Accountability

Section 4

NIST AI Risk Management Framework
AI Risk Management in Practice
CSF 2.0 Connection & Synthesis

SECTION 1

Professional Ethics & Stakeholder Analysis

Building the ethical foundation for cybersecurity decisions

Why Professional Ethics in Cybersecurity?

Cybersecurity professionals hold extraordinary power:

- Access to systems, data, and communications that others cannot see
- Ability to monitor, intercept, or disrupt digital activity
- Decisions that affect individuals, organizations, and public safety

With that power comes ethical responsibility:

- Decisions go beyond technical correctness — they carry moral weight
- Ethics provide guidance when law, policy, and tools fall short
- Professional codes create shared standards across the field

Professional Codes of Ethics

Core Frameworks

(ISC)² Code of Ethics

- Protect society, the common good, necessary public trust and confidence
- Act honorably, honestly, justly, responsibly, and legally
- Provide diligent and competent service to principals
- Advance and protect the profession

ACM Code of Ethics

- Contribute to society and human well-being
- Avoid harm — actively, not just passively
- Be honest and trustworthy
- Respect privacy and give proper credit

Shared Principles

What codes share in common:

- Honesty and integrity
- Avoiding harm
- Confidentiality of information
- Public interest beyond client interest
- Ongoing professional development

Competing Professional Responsibilities

Lower-Stakes Tensions

Duties to the Employer

- Protect the organization's systems and data
- Follow internal policies and directives
- Maintain confidentiality of organizational information
- Support business objectives and continuity

Duties to the Client / User

- Protect user data and privacy
- Operate only within authorized scope
- Disclose conflicts of interest

Higher-Stakes Tensions

Duties to the Profession

- Uphold ethical standards even when inconvenient
- Report misconduct through appropriate channels
- Maintain and improve competence

Duties to the Public

- Protect public safety and critical infrastructure
- Do not allow harm to others for organizational benefit
- Whistleblowing may be required when public safety is at stake

These duties can conflict — and that's where ethics gets hard.

Mini-Scenario: Where Do Your Loyalties Lie?

You are a security analyst. Your manager asks you to monitor the personal email of an employee suspected of leaking trade secrets — without the employee's knowledge and without any formal authorization. Your manager says, "Just do it quietly. No one will know."

Discussion Questions

- Which professional duties are in tension in this situation?
- What does your professional code of ethics tell you to do?
- What is the difference between what you are being asked to do and a legally authorized investigation?
- What would you actually do, and why?

Stakeholder Analysis in Cybersecurity Decisions

What is a stakeholder?

- Anyone who has an interest in or is affected by a decision or its outcomes

Why does it matter?

- Security decisions rarely affect only one person or group
- Ignoring stakeholders leads to incomplete analysis and harmful outcomes
- Identifying stakeholders helps us ask: "Who benefits? Who is harmed? Who has authority?"

Stakeholder categories in cybersecurity:

- Direct: the organization, the security team, individual employees under investigation
- Indirect: customers whose data is held, business partners, regulators
- Systemic: communities affected by security failures or over-monitoring
- Adversarial: threat actors who benefit from weak security

Stakeholder Map: The Insider Threat Case

The Flagged Employee

Privacy rights, presumption of innocence, career impact if wrongly accused, rights to due process

The Security Team

Professional duty to investigate threats; also duty not to harm innocent people; owns the final recommendation

The Organization

Protect intellectual property and systems; comply with law; minimize liability; maintain employee trust

Customers & Partners

Expect that their data is protected; harmed if a real threat goes unaddressed or privacy is violated

Legal / Compliance

Authorization requirements; labor law; data protection regulations; must be involved before action

The Public

Broader interest in organizations not using security tools as surveillance tools against legitimate activity

SECTION 2

Privacy, Authorization & The Case

Before the AI flagged anyone — what rules should govern?

Privacy & Authorization: Core Concepts

Privacy Foundations

Privacy in the Workplace

- Employees do not surrender all privacy rights at work
- Monitoring scope must be disclosed, proportionate, and lawful
- Different standards apply to company-owned vs. personal devices
- Privacy expectations vary by jurisdiction and organization policy

Key Legal Frameworks (U.S.)

- Electronic Communications Privacy Act (ECPA)
- Computer Fraud and Abuse Act (CFAA)
- State-level data protection laws (e.g., CCPA)
- Sector-specific rules (HIPAA, FERPA, GLBA)

Authorization Principles

Authorization

- Authorization is explicit permission to take a specific action
- Security tools do not grant unlimited surveillance authority
- "I can access it technically" ≠ "I am authorized to access it"

Proportionality

- Monitoring response must match the severity of the suspected threat
- Less invasive options should be exhausted before escalating
- Over-monitoring can violate rights even if technically lawful

Rule of Thumb:

- "What are we authorized to do, and is this the minimum necessary action?"

⚠ ALERT: AI INSIDER THREAT FLAG

Applied Case: AI Flags an Insider Threat

The AI behavioral monitoring system has generated a high-confidence insider threat alert. The flagged individual is a cybersecurity student employed in your organization's IT department. The system detected a cluster of anomalous behaviors over the past two weeks.

Flagged Behaviors

Unusual login times	Accessing systems at 2–4 AM on multiple nights
Privacy-tool use	VPN and Tor browser detected on corporate network
Virtual machines	Multiple VMs created on personal machine tied to corporate account
Network scanning	Nmap-style port scan activity detected on the internal network
Downloading resources	Large volume of cybersecurity tools, documentation, and exploit databases downloaded

SECTION 3

AI Threat Detection, Bias & Human Oversight

Understanding what the machine sees — and what it misses

How AI-Assisted Threat Detection Works

Behavioral Baseline Analysis

- System establishes a "normal" behavioral profile for each user over time
- Flags deviations from baseline: unusual times, volumes, destinations, tools

Machine Learning Pattern Matching

- Model is trained on data from known malicious insiders (historical cases)
- Looks for clusters of behaviors that resemble past threat patterns

Risk Scoring

- Individual behaviors generate risk scores; scores aggregate into an alert
- High-confidence alerts trigger human review — or automated response

Key Limitation:

- The system sees behavior — not intent, not context, not legitimate purpose

Analyzing the Evidence: The Case

The Machine's View

What the System SEES

- Unusual login times → anomaly flagged
- Privacy tools (VPN, Tor) → policy violation potential
- Virtual machines → unauthorized environment
- Network scanning → reconnaissance behavior
- Cybersecurity tool downloads → suspicious content

System's Implicit Assumptions

- Normal employees don't do these things
- These behaviors together = malicious intent
- Past malicious insiders showed similar patterns

The Human Context

What the System CANNOT SEE

- Student is enrolled in cybersecurity coursework
- Late-night logins = studying after a second job
- Privacy tools = coursework on privacy technology
- VMs = required lab environments for class
- Network scanning = authorized class exercise
- Tool downloads = required course materials

Missing Information

- Has the employee disclosed their coursework?
- Is there an acceptable use policy for lab work?
- What does the employee's manager know?

False Positives: What Are the Costs?

Human Harm

False Positive Defined

- System flags a person as a threat when they are not one
- In this case: a student doing legitimate coursework

Costs to the Individual

- Wrongful termination or disciplinary action
- Reputational damage — "suspected of insider threat"
- Psychological harm, loss of trust in employer
- Legal exposure even if ultimately cleared

Organizational & Systemic Harm

Costs to the Organization

- Legal liability for wrongful action against employee
- Loss of a skilled cybersecurity professional
- Erosion of employee trust in security programs
- Regulatory and HR consequences

Systemic Costs

- Alert fatigue — analysts stop trusting the system
- Chilling effect on legitimate security education
- Undermines confidence in AI-assisted security programs

False Negatives: The Other Side of the Equation

The Risk of Missing a Real Threat

False Negative Defined

- System fails to flag a genuine insider threat
- The threat goes undetected — and unaddressed

What can happen?

- Data exfiltration: sensitive records, IP, or credentials stolen
- Sabotage of critical systems or data integrity
- Cascading harm to customers, partners, or the public
- Regulatory fines, civil liability, reputational destruction

Why the Tradeoff Is Ethical, Not Technical

The Dilemma

- Reducing false positives (more lenient threshold) → more false negatives
- Reducing false negatives (stricter threshold) → more false positives
- There is no perfect system — only tradeoffs

This is why human judgment matters:

- The AI generates an alert; humans must contextualize it
- The organization's risk tolerance shapes the threshold
- Ethics guides what we do with that threshold decision

Bias and Assumptions in AI Threat Detection

Where does bias enter?

- Training data: if known malicious insiders were disproportionately from certain groups, the model may over-flag those groups
- Baseline assumption: "normal" is defined by the majority — unusual behavior is flagged even when it reflects legitimate difference
- Label bias: who gets called an insider threat affects what the model learns

Assumptions embedded in this case's alert:

- Working late = suspicious (ignores second jobs, caregiving, different work styles)
- Using privacy tools = hiding something (ignores privacy education, personal preference)
- Network scanning = attack reconnaissance (ignores authorized coursework)

Key principle:

- The behaviors that look suspicious are often the same behaviors that characterize cybersecurity professionals learning their craft.

Human Oversight: Why the Human Must Stay in the Loop

The automation bias risk

- Automation bias: humans over-trust algorithmic outputs and under-apply their own judgment
- "The AI said so" becomes a substitute for critical thinking

Automated action is not the same as automated decision

- Alert generation → automated (appropriate)
- Preliminary investigation → human-assisted (appropriate)
- Access revocation → requires human authorization (required)
- Termination, legal action, criminal referral → human decision with oversight (mandatory)

Human oversight functions:

- Context that AI cannot access (HR records, manager knowledge, employee disclosures)
- Proportionality judgment (is this response commensurate with the risk?)
- Accountability — only humans can be held professionally and legally accountable

Should Automated Detection Lead to Automated Action?

The AI system is configured with an optional "auto-response" mode. When an insider threat alert reaches a 90% confidence level, the system can automatically revoke network access, disable login credentials, and notify HR — without any human review. Your organization is considering enabling this feature.

Discussion Questions

- What could go wrong if auto-response is enabled? What are the benefits?
- Who is accountable if a false positive leads to wrongful termination under auto-response?
- Is a 90% confidence threshold "good enough" for automated action? Why or why not?
- What human review process should occur before any action is taken — regardless of confidence level?
- What would the NIST AI RMF (Govern/Map/Measure/Manage) say about this design choice?

Professional Accountability: Who Owns the Decision?

The security analyst who acts on the alert

- Professional duty to investigate, not simply execute
- Code of ethics applies to how they treat the flagged employee

The security manager who authorizes action

- Bears responsibility for decisions made under their authority
- Cannot delegate ethics to the algorithm

The organization that deploys the AI

- Responsible for configuring appropriate thresholds and review processes
- Liable for harm caused by the system operating as designed

Critical principle:

- "The AI flagged it" is not a defense. Professionals remain accountable for decisions made with AI tools — just as a driver remains accountable when GPS leads them the wrong way.

SECTION 4

NIST AI Risk Management Framework & CSF 2.0

Governance tools for AI-enabled cybersecurity

NIST AI Risk Management Framework: Core Functions

GOVERN

Who is responsible?
What policies and oversight should exist?
Establish culture, accountability, and authority for AI risk.

MAP

What is happening?
Who could be affected?
What could go wrong?
Identify context, stakeholders, and risks.

MEASURE

How do we know if the system performs appropriately?
Quantify, test, evaluate, and monitor AI risk.

MANAGE

What should the organization do about identified risks?
Prioritize, respond, and treat AI risks across the lifecycle.

Applying GOVERN: Who Is Responsible?

Policy & Oversight

What policy governs how AI insider threat alerts are handled? Who approved deploying this system?

Roles & Authority

Who is authorized to take action on an alert? Who must approve each escalation step?

Accountability Structures

If the system causes harm, who is responsible? Is there a documented chain of accountability?

Ethical Culture

Does the organization have a culture where security professionals can push back on the AI's outputs without fear?

Applying MAP: What Could Go Wrong?

What is happening?

An AI system monitors employee behavior and generates insider threat alerts with a confidence score and recommended action.

Who could be affected?

Flagged employees, the security team, HR, legal, the broader workforce watching how cases are handled, customers whose data is at stake.

What could go wrong?

False positives harm innocent employees. False negatives allow real threats. Bias systematically disadvantages certain groups. Auto-response causes irreversible harm.

What context matters?

The flagged person's role, their disclosures, the specific behaviors and their plausible explanations, legal authorization for each investigative step.

Applying MEASURE: How Do We Know If It's Working?

Performance metrics for the AI system:

- False positive rate: what % of alerts are for innocent individuals?
- False negative rate: what % of real threats does the system miss?
- Time to human review: how long from alert to human decision?
- Demographic analysis: are any groups disproportionately flagged?

Human decision quality metrics:

- Are analysts applying consistent criteria to similar alerts?
- Are escalations documented with clear reasoning?
- Do outcomes reflect the intent of the policy?

Key principle: You cannot manage what you do not measure.

- Measuring only "did the system flag a threat?" is insufficient — you must also measure what the system got wrong and why.

Applying MANAGE: What Should the Organization Do?

Immediate (this alert)

Pause automated action. Human review with full context. Contact manager and HR. Interview the employee before any action.

Short-Term

Establish clear escalation policy. Define what authorization is required at each step. Document every decision and its justification.

Medium-Term

Audit the AI system's false positive rate. Test for demographic bias. Adjust thresholds based on findings. Train analysts to contextualize alerts.

Long-Term

Build a review board for AI-generated security decisions. Create a formal disclosure process for employees doing security coursework. Align policy with professional codes of ethics.

Connecting to NIST Cybersecurity Framework 2.0

Framework Overview

What is CSF 2.0?

- A framework for managing cybersecurity risk across an organization
- Updated in 2024 with GOVERN as a new core function
- Designed to work alongside other frameworks including AI RMF

CSF 2.0 Core Functions:

- GOVERN → Organizational context and risk management strategy
- IDENTIFY → Asset management, risk assessment
- PROTECT → Access control, training, data security
- DETECT → Anomaly detection (where our AI system sits)
- RESPOND → Incident response planning and execution
- RECOVER → Recovery planning and communications

Applied to Our Case

Today's Course Connection:

- Our AI insider threat system lives in DETECT
- But the ethical issues touch every function

Example:

- GOVERN: Does policy authorize what the AI does?
- IDENTIFY: Have we identified insider threat as a risk — and false positives as a risk?
- PROTECT: Do access controls limit what the AI can see?
- DETECT: Is the AI's detection accurate and fair?
- RESPOND: What do we do when a human reviews the alert?

Takeaway: Cybersecurity decisions exist within an organization, not just inside the security team.

Full Case Debrief: What Should Your Team Do?

Your team has received the AI insider threat alert. You have reviewed the flagged behaviors. You now know the employee is a cybersecurity student. Your manager wants a recommendation before the end of the day.

Discussion Questions

- GOVERN: Who in your organization has the authority to authorize the next investigative step? Who should NOT make this decision alone?
- MAP: Who are all the stakeholders? What additional information do you need before making a recommendation?
- MEASURE: What would you need to know about the AI system's accuracy before trusting this alert?
- MANAGE: What is your team's recommendation? What action (if any) is proportionate, authorized, and ethically defensible?
- Accountability: Who signs off on whatever decision is made, and what documentation should exist?

AI Risk Management: Principles to Carry Forward

1. AI is a tool, not a decision-maker.

- Alerts, flags, and scores require human interpretation before any consequential action.

2. The model sees behavior — not intent.

- Context that the system cannot access is exactly the context human judgment must provide.

3. Bias is built in if you don't build it out.

- All AI systems reflect the assumptions in their training data. Those assumptions must be examined.

4. Authorization is not optional.

- "The AI said so" does not constitute authorization to investigate, discipline, or terminate.

5. Professionals remain accountable.

- Deploying AI does not transfer ethical or legal accountability to the algorithm.

Four Statements You Should Be Able to Explain

If you can explain why these are true, you understand the ethical foundation of this course.

Technical capability does not equal authorization.

Legal does not automatically mean ethical.

AI-generated does not automatically mean accurate.

Professional judgment cannot simply be outsourced to a tool.

Exit Ticket — turn in before you leave:

- Choose one of the four statements. Write 3–4 sentences that explain why it is true using a specific example from today's case or discussion.
- Be specific. "Because AI is not always right" is not enough. Show that you understand what the statement means and why it matters for cybersecurity professionals.

Unpacking the Four Statements

Technical capability does not equal authorization.

What it means:

- Access to a system, data, or account does not mean you are permitted to use that access.
- Authorization requires explicit permission within a defined scope, granted by someone with the authority to grant it.

From today's case:

- The security team can technically read the employee's login logs, examine their VMs, and intercept their traffic. None of that is authorized without proper approval and legal review.
- "I can access it" is the starting point of a question, not the answer to one.

Legal does not automatically mean ethical.

What it means:

- Law sets a floor, not a ceiling. An action can be legally permissible and still violate professional ethics, employee rights, or the public interest.
- Professional codes of ethics exist precisely because law alone is insufficient to govern complex decisions.

From today's case:

- Monitoring an employee's network activity may be legal under an employer's acceptable use policy. Whether it is proportionate, fair, and consistent with the employee's dignity is an ethical question the law does not answer.

AI-generated does not automatically mean accurate.

What it means:

- AI systems produce outputs based on patterns and probabilities, not truth. A high-confidence alert is not a finding of fact.
- All AI systems have error rates. Without knowing the false positive and false negative rates, a single alert tells you very little.

From today's case:

- The AI flagged a cybersecurity student doing legitimate coursework. Every behavior it cited had an innocent explanation the system could not see. "90% confidence" does not mean the alert is correct — it means the algorithm matched a pattern.

Professional judgment cannot simply be outsourced to a tool.

What it means:

- Ethical and professional responsibility requires a human being to reason through the decision — its context, its stakeholders, its proportionality, and its consequences.
- Automation bias — the tendency to defer to algorithmic outputs — does not transfer the professional's accountability. It just means they failed to exercise it.

From today's case:

- "The AI flagged it" is not a defense if the action harms an innocent person. The analyst, the manager, and the organization all bear responsibility for what happens next — regardless of what the system recommended.

Class Meeting 2: Key Takeaways

Professional Ethics

Codes of ethics guide decisions under pressure. Competing duties are real — the public interest does not disappear because your employer asks you to ignore it.

Stakeholder Analysis

Always ask who else is affected. The flagged employee, their family, other workers watching the outcome, and customers are all stakeholders the AI does not see.

Privacy & Authorization

"I can access it" is not authorization. Every investigative step requires explicit authority proportionate to the suspected threat.

AI Limitations

AI sees patterns, not people. False positives and false negatives both have real costs. Bias in, bias out.

Human Oversight

Automation bias is a real risk. Humans must remain in the decision loop for all consequential outcomes — and must actually exercise judgment, not rubber-stamp the algorithm.

NIST AI RMF

GOVERN → MAP → MEASURE → MANAGE. These four functions provide a systematic approach to AI risk that complements your ethical reasoning.

Resources & Looking Ahead

NIST AI RMF https://www.nist.gov/itl/ai-risk-management-framework	Full framework documentation and supporting resources
NIST CSF 2.0 https://www.nist.gov/cyberframework	Cybersecurity Framework 2.0 — full text and quick-start guides
AI RMF Explainer Video https://www.nist.gov/video/introduction-nist-ai-risk-management-framework-ai-rmf-10-explainer-video	Optional: 10-minute NIST introduction video
(ISC)² Code of Ethics https://www.isc2.org/Ethics	Canonical professional code for security certification holders
ACM Code of Ethics https://www.acm.org/code-of-ethics	Broader computing ethics framework applicable to all tech professionals

Questions?

The goal of this course is not to make you suspicious of AI — it is to make you the kind of professional who asks the right questions before acting on it.