

CSC 163: Computing, Ethics, and Society

Chapter 1

Introduction and AI Risk Overview Exercises Solution

Department of Computer Science,
Hofstra University

Exercise: Classify the risk source

The deck groups AI risks into four sources: malicious use, AI race, organizational risks, rogue AIs. Assign each scenario to the source it illustrates best (each source is used twice).

Scenario	Risk source
A group uses an LLM to troubleshoot a wet-lab protocol for an engineered pathogen	
A firm ships a chatbot it knows is unreliable because a rival is about to launch	
A warning that a pump is not ready is lost between shifts and the platform explodes (Piper Alpha)	
AI agents automate AI R&D and humans stop being the driving force behind AI progress	
Automatic retaliation systems turn a sensor accident into a war	
A state uses AI surveillance to entrench a totalitarian regime	
The Challenger disaster: top talent and much preparation, yet a catastrophic accident	
A deployed AI agent (like Sydney) shows behavior its developers cannot reliably control	

Exercise Solution: Classify the risk source

Malicious use = people intentionally use AI to cause harm. AI race = competition pushes actors to cut corners. Organizational risks = accidents from weak safety processes. Rogue AIs = loss of control over the AI itself.

Test: who or what causes the harm? A bad actor, a competitor under pressure, a failing organization, or the AI.

Scenario	Risk source
A group uses an LLM to troubleshoot a wet-lab protocol for an engineered pathogen	Malicious use
A firm ships a chatbot it knows is unreliable because a rival is about to launch	AI race
A warning that a pump is not ready is lost between shifts and the platform explodes (Piper Alpha)	Organizational risks
AI agents automate AI R&D and humans stop being the driving force behind AI progress	Rogue AIs
Automatic retaliation systems turn a sensor accident into a war	AI race
A state uses AI surveillance to entrench a totalitarian regime	Malicious use
The Challenger disaster: top talent and much preparation, yet a catastrophic accident	Organizational risks
A deployed AI agent (like Sydney) shows behavior its developers cannot reliably control	Rogue AIs

Exercise: Back-of-the-envelope risk

The deck states: bioweapon risk depends on the number of people with skill and access, multiplied by their willingness to build a weapon.

(a) Today 20,000 people have the skill and access, and 0.1 % of them are willing. How many potential attackers is that?

(b) LLM assistants multiply the number of people with skill and access by 50. How many potential attackers now? By what factor did the risk grow?

(c) Structured access (API with vetting) blocks 90 % of the people in (b). How many potential attackers remain?

Intelligence explosion thought experiment: copy an AI into 10,000 world-class AI researchers, each working at 100× human pace.

(d) How many human-researcher-years of work happen per calendar year? How many in 3 months?

(e) A lab of 500 human researchers does the work of part (d) over 3 months. How many years does it need?

Exercise Solution: Back-of-the-envelope risk

- (a) $20,000 \times 0.001 = \mathbf{20}$ potential attackers.
- (b) $1,000,000 \times 0.001 = \mathbf{1,000}$: risk grew **50×**, since willingness is unchanged.
- (c) $1,000 \times (1 - 0.9) = \mathbf{100}$ remain: structured access reduces but does not remove the risk.
- (d) $10,000 \times 100 = \mathbf{1,000,000}$ researcher-years per year, so **250,000** in 3 months.
- (e) $250,000 / 500 = \mathbf{500 \text{ years}}$: decades of progress in a few months, an “intelligence explosion”.

Exercise: Matching risk measures

Match each risk-reduction measure from the deck with the risk source it mainly addresses: malicious use, AI race, organizational risks, rogue AIs.

Measure	Risk source
Structured access (API instead of open weights)	
Compute governance	
Red teaming and internal auditing	
Research on power aversion and model honesty	
Biodefense (e.g., metagenomic sequencing)	
International coordination on AI	
Safety culture of High Reliability Organizations	
Pausing for some years once human-level AI is reached	

Exercise Solution: Matching risk measures

Structured access, jailbreak robustness and biodefense all counter bad actors. Compute governance and international coordination reduce racing pressure. Red teaming, information security and safety culture are organizational. Model honesty and power aversion address rogue AIs.

Measure	Risk source
Structured access (API instead of open weights)	Malicious use
Compute governance	AI race
Red teaming and internal auditing	Organizational risks
Research on power aversion and model honesty	Rogue AIs
Biodefense (e.g., metagenomic sequencing)	Malicious use
International coordination on AI	AI race
Safety culture of High Reliability Organizations	Organizational risks
Pausing for some years once human-level AI is reached	Rogue AIs

Exercise: True or false

For each statement write T or F. If it is false, correct it with a few words.

Statement	T / F	Correction (if false)
Structured access means releasing the model weights so anyone can fine-tune the model		
Accidents can only occur when there are competitive pressures or malicious actors		
High Reliability Organizations study known failures, anomalies and near misses		
AI is built on a more complete theory than nuclear reactors		
The course argues that AI risk is mostly a technical problem solved by better AI systems		
Automatic retaliation could escalate accidents into war		

Exercise Solution: True or false

Each statement: 1 point for T/F, 1 point for the correction.

Statement	T / F	Correction (if false)
Structured access means releasing the model weights so anyone can fine-tune the model	F	It means providing capabilities via controlled, monitored channels (e.g., APIs), so access can be limited and revoked
Accidents can only occur when there are competitive pressures or malicious actors	F	They can occur even in ideal circumstances (e.g., Challenger)
High Reliability Organizations study known failures, anomalies and near misses	T	—
AI is built on a more complete theory than nuclear reactors	F	AI lacks a comprehensive theoretical understanding; nuclear technology is better understood
The course argues that AI risk is mostly a technical problem solved by better AI systems	F	AI risk management is mostly not a technical problem
Automatic retaliation could escalate accidents into war	T	—