

CSC 163: Computing, Ethics, and Society

Chapter 3

Single Agent Safety Exercises Solution

Department of Computer Science,
Hofstra University

Exercise: Four research areas

ML safety has four research areas: robustness (withstand hazards), monitoring (identify hazards), control (reduce inherent model hazards), systemic safety (reduce risks beyond individual models). Place each technique in one area.

Technique	Area
Adversarial training	
Cleansing a poisoned (Trojan) model	
Anomaly detection (out-of-distribution inputs)	
Mechanistic interpretability	
Representation control	
Machine unlearning of hazardous knowledge	
Watermarking AI-generated text	
ML intrusion detection for cyberdefense	

Exercise Solution: Four research areas

Robustness: attacks and defenses of the model itself. Monitoring: finding out what a model can do or is doing. Control: changing the model's inner behavior. Systemic safety: protecting the world around the model.

Technique	Area
Adversarial training	Robustness
Cleansing a poisoned (Trojan) model	Robustness
Anomaly detection (out-of-distribution inputs)	Monitoring
Mechanistic interpretability	Monitoring
Representation control	Control
Machine unlearning of hazardous knowledge	Control
Watermarking AI-generated text	Systemic safety
ML intrusion detection for cyberdefense	Systemic safety

Exercise: Anomaly detection scores

An anomaly detector gives each input a score; an input is flagged if its score is greater than the threshold $\tau = 0.5$. Four inputs are truly out-of-distribution (OOD).

(a) Fill in whether each input is flagged and its outcome (TP, FP, FN, TN; OOD is the positive class).

(b) Compute precision = $TP/(TP+FP)$ and recall = $TP/(TP+FN)$ at $\tau = 0.5$.

(c) Lower the threshold to $\tau = 0.3$. Recompute precision and recall. What is the trade-off?

Input	True class	Score	Flagged at 0.5?	Outcome
A	In-dist	0.05		
B	OOD	0.92		
C	In-dist	0.40		
D	OOD	0.35		
E	In-dist	0.70		
F	OOD	0.88		
G	In-dist	0.10		
H	OOD	0.65		

Exercise Solution: Anomaly detection scores

- (a) Table: A, C, G are true negatives; B, F, H are true positives; D is missed (FN); E is a false alarm (FP).
- (b) $TP = 3$, $FP = 1$, $FN = 1$: precision = $3/4 = 0.75$, recall = $3/4 = 0.75$.
- (c) At $\tau = 0.3$ flagged: B, C, D, E, F, H; $TP = 4$, $FP = 2$, $FN = 0$: precision = $4/6 \approx 0.67$, recall = **1.0**.
- A lower threshold catches every anomaly (higher recall) at the price of more false alarms (lower precision).

Input	True class	Score	Flagged at 0.5?	Outcome
A	In-dist	0.05	No	TN
B	OOD	0.92	Yes	TP
C	In-dist	0.40	No	TN
D	OOD	0.35	No	FN
E	In-dist	0.70	Yes	FP
F	OOD	0.88	Yes	TP
G	In-dist	0.10	No	TN
H	OOD	0.65	Yes	TP

Exercise: Benchmark desiderata

A good benchmark has: high ceiling, automatic evaluability, ease of setting up, reproducibility, clear downstream implications, an interpretable metric, usefulness for hill climbing. Which desideratum does each benchmark violate most clearly?

Benchmark flaw	Desideratum violated
Every answer must be graded by a human	
Models score 99 % within a year; no headroom is left	
Scores change from day to day with no change to the model	
It requires specialized drivers and complicated software to run	
The score is only pass/fail, with no partial progress	
Score is in nats; stakeholders cannot interpret it	
Improving the score says nothing about useful real-world tasks	

Exercise Solution: Benchmark desiderata

Each flaw maps to one desideratum; hill climbing needs a smooth score, not a pass/fail indicator.

Benchmark flaw	Desideratum violated
Every answer must be graded by a human	Automatic evaluability
Models score 99 % within a year; no headroom is left	High ceiling
Scores change from day to day with no change to the model	Reproducible
It requires specialized drivers and complicated software to run	Ease of setting up
The score is only pass/fail, with no partial progress	Useful for hill climbing
Score is in nats; stakeholders cannot interpret it	Interpretable metric
Improving the score says nothing about useful real-world tasks	Clear downstream implications

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)
Adversarial training fully solves adversarial robustness at no cost		
An adversarial example crafted for one model can attack other models (transferability)		
A Trojan needs a large fraction (e.g., 30 %) of the training data to be poisoned		
Mechanistic interpretability is a top-down approach; representation engineering is bottom-up		
Machine unlearning needs retraining the model from scratch		
A more intelligent model is automatically a safer model		

Exercise Solution: True or false

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

Statement	T / F	Correction (if false)
Adversarial training fully solves adversarial robustness at no cost	F	It can cut clean accuracy by 10 %+, and the problem is still hard
An adversarial example crafted for one model can attack other models (transferability)	T	—
A Trojan needs a large fraction (e.g., 30 %) of the training data to be poisoned	F	A small fraction (e.g., 0.05 %) can be enough
Mechanistic interpretability is a top-down approach; representation engineering is bottom-up	F	Reversed: mechanistic is bottom-up, representation engineering is top-down
Machine unlearning needs retraining the model from scratch	F	It selectively removes knowledge without retraining from scratch
A more intelligent model is automatically a safer model	F	Intelligence cuts both ways; honesty, justice, kindness are not guaranteed