

CSC 163: Computing, Ethics, and Society

Chapter 4

Safety Engineering Exercises Solution

Department of Computer Science,
Hofstra University

Exercise: Nines of reliability

An autonomous vehicle makes a decision every time it must react. Its reliability p is the probability that a single decision is correct, and $k = -\log_{10}(1 - p)$ is its number of nines. On average a mistake occurs at about decision number $1/(1 - p)$.

(a) Fill in the table.

(b) A trip needs 1000 decisions and $p = 0.999$. What is the probability that the whole trip has no mistake?

(c) By what factor does the expected time to a mistake grow when p improves from 0.4 to 0.8? From 0.99 to 0.999?

p	k (nines)	Expected decisions to first mistake
0.9		
0.999		
0.99999		
	6	

Exercise Solution: Nines of reliability

- (a) $k = -\log_{10}(1 - p)$; expected decisions = $1/(1 - p)$. The last row is read backwards: $k = 6$ means $p = 0.999999$.
- (b) $P(\text{no mistake}) = 0.999^{1000} \approx \mathbf{0.37}$, so about **63 %** of such trips contain at least one mistake.
- (c) $0.4 \rightarrow 0.8$: $1/0.6 \rightarrow 1/0.2$, only **$\times 3$** . $0.99 \rightarrow 0.999$: $100 \rightarrow 1000$, **$\times 10$** .
- **Each added nine multiplies the expected time to failure by 10**; the gain is not linear in p . Exposure matters: more decisions, more chances to fail.

p	k (nines)	Expected decisions to first mistake
0.9	1	10
0.999	3	1,000
0.99999	5	100,000
0.999999	6	1,000,000

Exercise: Risk equation

Classic risk equation: $\text{Risk} = \sum P(\text{hazard}) \times \text{severity}(\text{hazard})$. Three hazards of an AI deployment are listed below (severity in millions of dollars).

(a) Compute the expected annual risk of each hazard and the total, and name the hazard with the highest risk.

(b) A new safeguard cuts the leak probability by 75 %; another halves the severity of a deepfake fraud (lower vulnerability). New total risk? Which safeguard reduces risk more?

Hazard	P (per year)	Severity (M\$)	Risk (M\$ / year)
Chatbot data leak	0.2	5	
Deepfake fraud wave	0.05	40	
Critical-infrastructure failure	0.002	1500	
Total			

Exercise Solution: Risk equation

(a) Risk = $P \times \text{severity}$: $0.2 \times 5 = 1.0$; $0.05 \times 40 = 2.0$; $0.002 \times 1500 = 3.0$. Total **6.0** M\$/year. The rarest hazard has the highest risk: **critical-infrastructure failure** (a tail event).

(b) Leak: $0.05 \times 5 = 0.25$ (-0.75). Deepfake: $0.05 \times 20 = 1.0$ (-1.0). New total = $0.25 + 1.0 + 3.0 = 4.25$ M\$/year. The **deepfake safeguard** reduces risk more (1.0 vs 0.75).

Ranking by probability alone would pick the wrong hazard: risk needs both factors.

Hazard	P (per year)	Severity (M\$)	Risk (M\$ / year)
Chatbot data leak	0.2	5	1.0
Deepfake fraud wave	0.05	40	2.0
Critical-infrastructure failure	0.002	1500	3.0
Total			6.0

Exercise: Safe design principles

Match each scenario with ONE of the eight safe design principles: redundancy, separation of duties, least privilege, fail-safe, antifragility, negative feedback, transparency, defense in depth. (Each principle is used at most once; one principle is not used.)

Scenario	Principle
A car has two independent braking systems	
An AI acts only if its confidence is above a threshold	
One employee approves a payment, a different one executes it	
A contractor AI gets read access only to files needed for its task	
Pilots are trained to activate and override the autopilot	
Mask, distancing and hand washing together against a virus	
Muscles grow stronger after being strained by exercise	

Exercise Solution: Safe design principles

Eight principles: redundancy, separation of duties, least privilege, fail-safe, antifragility, negative feedback, transparency, defense in depth. **Negative feedback** is the unused one.

Redundancy = several copies of the same function; defense in depth = several different protective layers.

Scenario	Principle
A car has two independent braking systems	Redundancy
An AI acts only if its confidence is above a threshold	Fail-safe
One employee approves a payment, a different one executes it	Separation of duties
A contractor AI gets read access only to files needed for its task	Least privilege
Pilots are trained to activate and override the autopilot	Transparency
Mask, distancing and hand washing together against a virus	Defense in depth
Muscles grow stronger after being strained by exercise	Antifragility

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)
Swiss cheese model: safety improves by making holes larger and using fewer slices		
Fault tree analysis starts from a single component and predicts all outcomes it might cause		
Adding one nine of reliability multiplies the expected time to failure by ten		
A lottery with EV = +\$4.985 (99.9 % win \$15, 0.1 % lose \$10,000) is safe even if probabilities are slightly off		
Every tail event is a black swan		
Defense in depth can introduce the risk of layers interacting in unexpected ways		

Exercise Solution: True or false

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

Key idea: tail events render averages inadequate.

Statement	T / F	Correction (if false)
Swiss cheese model: safety improves by making holes larger and using fewer slices	F	Reduce the size and number of holes, increase the number of slices
Fault tree analysis starts from a single component and predicts all outcomes it might cause	F	It starts from a specific negative outcome and works backward to its causes
Adding one nine of reliability multiplies the expected time to failure by ten	T	—
A lottery with EV = +\$4.985 (99.9 % win \$15, 0.1 % lose \$10,000) is safe even if probabilities are slightly off	F	At 99.7 % / 0.3 % the EV is -\$15.045: EV is very sensitive to probabilities
Every tail event is a black swan	F	Black swans are only the largely unanticipated tail events (unknown unknowns)
Defense in depth can introduce the risk of layers interacting in unexpected ways	T	—