

CSC 163: Computing, Ethics, and Society

Chapter 6

Beneficial AI and Machine Ethics Exercises Solution

Department of Computer Science,
Hofstra University

Exercise: Fairness metrics

A risk tool flags people as high risk. Confusion counts (positive = flagged high risk):

Group A: TP 30, FP 10, FN 10, TN 50.

Group B: TP 45, FP 15, FN 10, TN 30.

(a) Fill in the table. Equal across groups?

(b) Which of the three definitions (statistical parity, equalized odds, calibration) does the tool satisfy?

(c) How does this relate to the COMPAS dispute?

Metric	A	B	Equal?
Positive rate			
False positive rate			
False negative rate			
Precision			

Exercise Solution: Fairness metrics

- (a) Positive rate = $(TP+FP)/N$; FPR = $FP/(FP+TN)$; FNR = $FN/(FN+TP)$; precision = $TP/(TP+FP)$.
- (b) Only **calibration**: precision is 75 % in both groups. Positive rate (40 % vs 60 %), FPR and FNR differ, so statistical parity and equalized odds fail.
- (c) COMPAS: claim of bias (unequal error rates) vs counterclaim of fairness (calibrated). Both can be true, since the definitions are usually incompatible (impossibility theorem).

Metric	A	B	Equal?
Positive rate	40 %	60 %	No
False positive rate	16.7 %	33.3 %	No
False negative rate	25 %	18.2 %	No
Precision	75 %	75 %	Yes

Exercise: Gini coefficient

Gini coefficient: $G = \sum_i \sum_j |x_i - x_j| / (2 n^2 \bar{x})$, the sum over all ordered pairs; 0 = perfect equality.

Five households' incomes (k\$): before automation 10, 10, 10, 10, 60. After automation (GPU owners capture the value) 5, 5, 5, 5, 80.

(a) Compute the mean and the Gini coefficient before and after.

(b) Name the market failure: (i) a lender exploits what the borrower does not know; (ii) one firm is big enough to set the price; (iii) a data center's pollution harms neighbors who are not its customers.

	Mean (k\$)	Gini
Before		
After		

Exercise Solution: Gini coefficient

- (a) Before: $\bar{x} = 20$; the sum over ordered pairs $= 2 \times 4 \times 50 = 400$; $G = 400 / (2 \times 25 \times 20) = \mathbf{0.40}$.
- After: $\bar{x} = 20$; sum $= 2 \times 4 \times 75 = 600$; $G = 600 / 1000 = \mathbf{0.60}$. Same mean, higher inequality.
- (b) (i) **information asymmetry**; (ii) **monopoly**; (iii) **externality**.

	Mean (k\$)	Gini
Before	20	0.40
After	20	0.60

Exercise: Moral uncertainty

An AI must choose among three options under moral uncertainty. Its credences: 60 % utilitarianism, 40 % deontology. Choice-worthiness of each option under each theory is in the table.

(a) Compute the expected choice-worthiness (ECW) of each option.

(b) Which option does "My Favorite Theory" choose? Which does "Maximize Expected Choice-Worthiness" choose?

(c) Which approach would let several stakeholder AIs negotiate and vote?

Option	Utilit. (0.6)	Deont. (0.4)	ECW
A	10	-20	
B	6	4	
C	0	0	

Exercise Solution: Moral uncertainty

- (a) $ECW = 0.6 \times U + 0.4 \times D$. A: $6 - 8 = -2.0$; B: $3.6 + 1.6 = 5.2$; C: 0 .
- (b) My Favorite Theory follows the highest-credence theory (utilitarianism): option **A** ($U = 10$). MEC picks the highest ECW: option **B**.
- (c) The **moral parliament**.

A violates the deontological constraint badly (-20): MEC accounts for that, My Favorite Theory ignores it.

Option	Utilit. (0.6)	Deont. (0.4)	ECW
A	10	-20	-2.0
B	6	4	5.2
C	0	0	0

Exercise: Preferences and wellbeing

Match each statement with the best concept: revealed preference, stated preference, idealized preference, wellbeing as pleasure, wellbeing as objective goods, wellbeing as preference satisfaction.

Statement	Concept
Alice binge-watches clickbait although she says she wants to stop (weakness of will)	
Alice ranks two AI outputs, and the ranking is used to train the model (DPO)	
What Alice would choose if she knew all the relevant facts	
What matters is the balance of pleasure over pain (hedonism)	
Learning is better for you than counting blades of grass, even if you desire the latter	
Wellbeing is having one's desires satisfied	
A recommender infers what users like from what they clicked	

Exercise Solution: Preferences and wellbeing

Preferences: revealed = what people choose; stated = what they say; idealized = what a fully informed person would choose. Weakness of will is a drawback of revealed preferences.

Statement	Concept
Alice binge-watches clickbait although she says she wants to stop (weakness of will)	Revealed preference
Alice ranks two AI outputs, and the ranking is used to train the model (DPO)	Stated preference
What Alice would choose if she knew all the relevant facts	Idealized preference
What matters is the balance of pleasure over pain (hedonism)	Wellbeing as pleasure
Learning is better for you than counting blades of grass, even if you desire the latter	Wellbeing as objective goods
Wellbeing is having one's desires satisfied	Wellbeing as preference satisfaction
A recommender infers what users like from what they clicked	Revealed preference