UNIVERSITÀ DI PAVIA

Department of Electrical, Computer
and Biomedical Engineering

# Steering AI: Legal Challenges and Ethical Standards from an Engineering Perspective

Thursday 18 April 2024

Museo della Tecnica Elettrica
Via Ferrata, 6
27100 Pavia

*With the patronage of:*

IL COLLEGIO
FONDAZIONE GHISLIERI

*With the contribution of:*

ReD OPEN
responsibility & design in open ecosystems

# Should AI Steer the Trolley?

## Steering AI: Legal Challenges and Ethical Standards from an Engineering Perspective

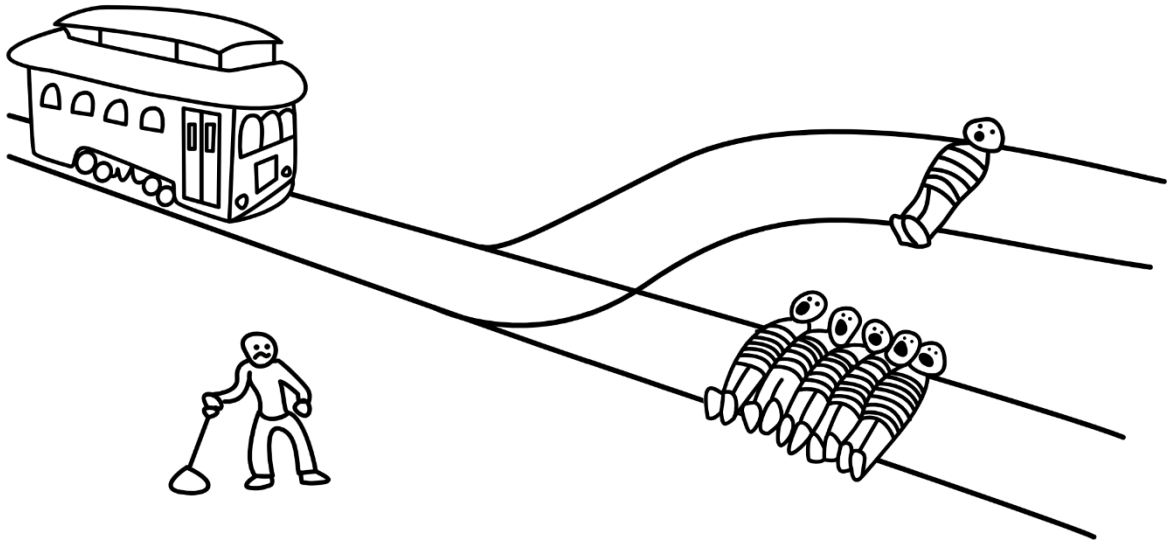
Antonio Barili
*Dept. of Electrical, Computer and Biomedical Engineering*
*Digital Forensics Laboratory*


R. FIGUEIRA
12
carris

NEAL.FUN

## Absurd Trolley Problems

Level 1: The Original



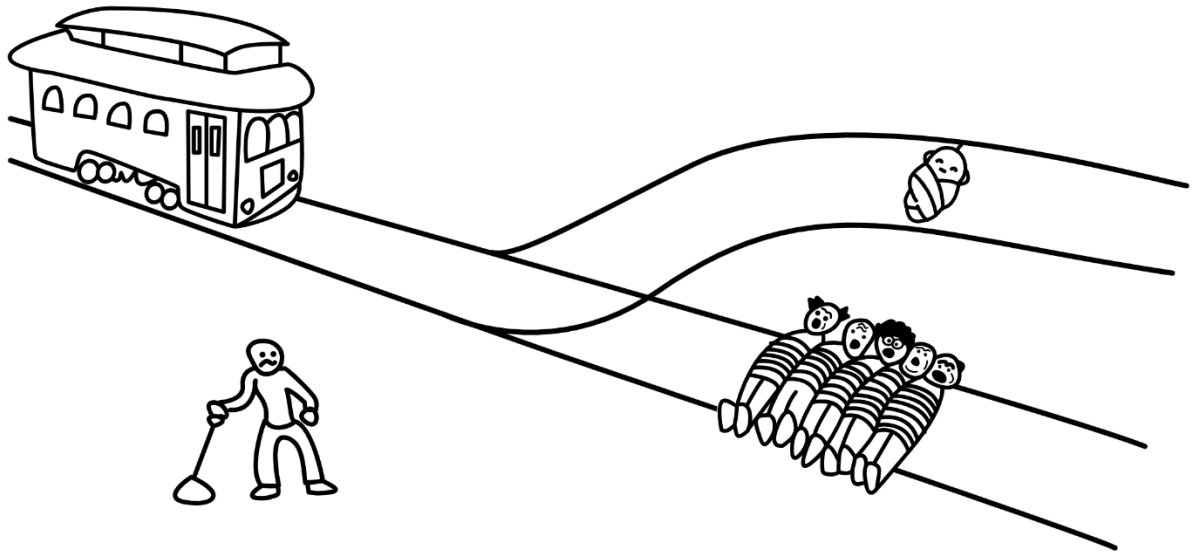
Oh no! A trolley is heading towards 5 people. You can pull the lever to divert it to the other track, killing 1 person instead. What do you do?

Pull the lever | Do nothing

NEAL.FUN

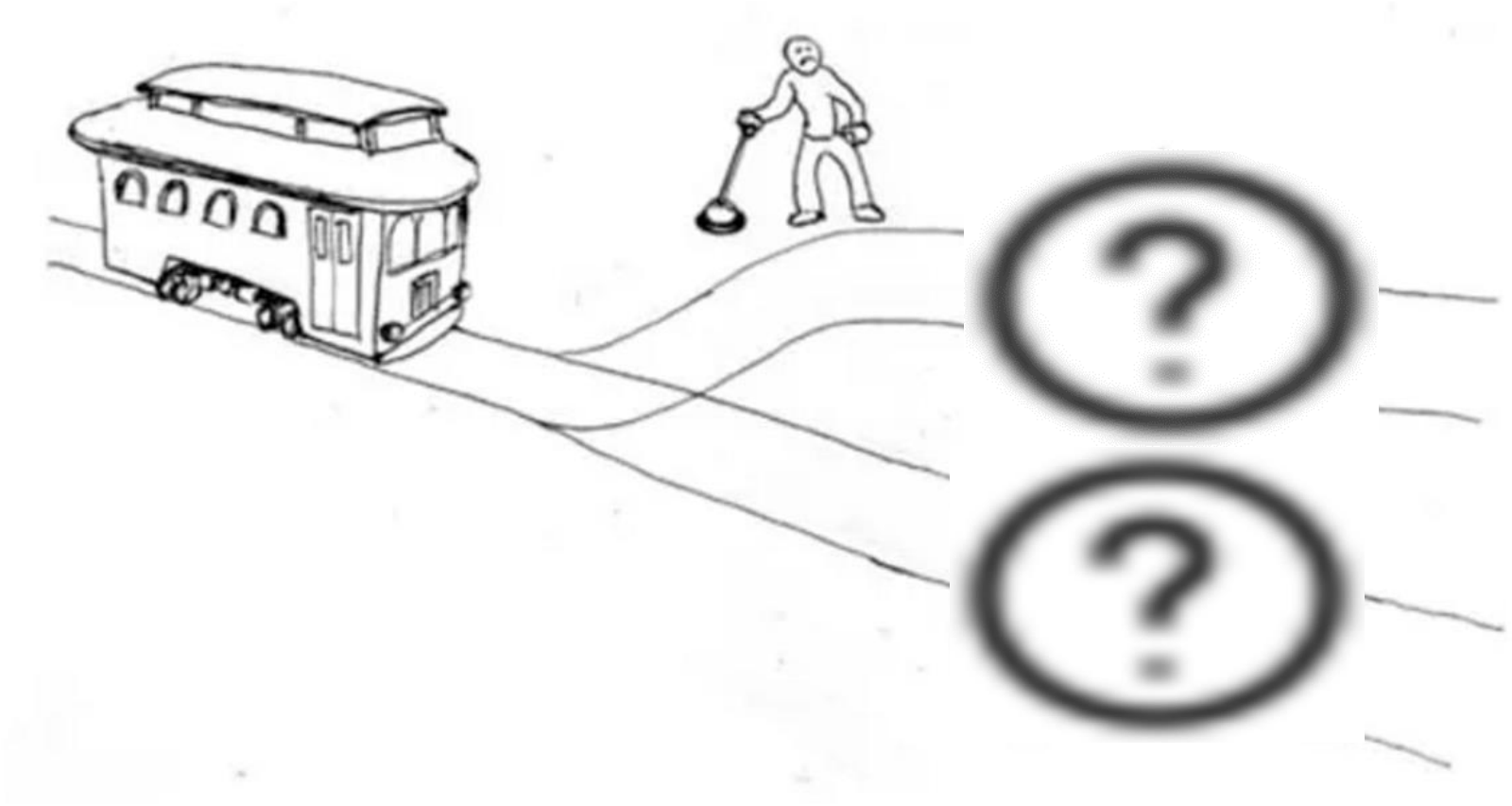
# Absurd Trolley Problems

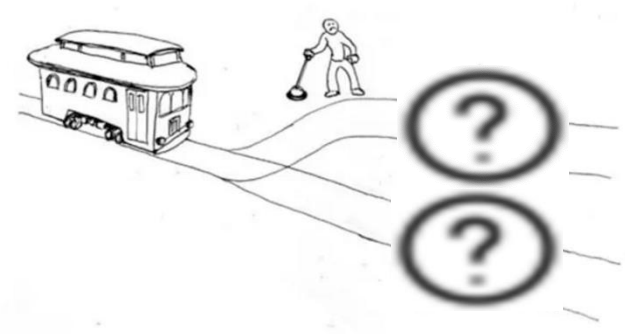
Level 15: Age



Oh no! A trolley is heading towards 5 elderly people. You can pull the lever to divert it to the other track, running over a baby instead. What do you do?

Pull the lever | Do nothing






1) Assess what's actually on each track

2) Decide how to steer the trolley

3) ... cross your fingers!

- A smartphone seized from a known drug smuggler stores about 100.000 pictures
- On average, 0.1% of them portray drug samples
- A forensic software implements an (AI-based) image classifier to tell drug pictures from non-relevant material

Nice, but ...

| Total Population (P+N) | Prevalence P/(P+N) | True Positive Rate TPR=TP/P | True Negative Rate (Specificity) TNR = TN/N |
|---|---|---|---|
| 100.000 | 0,001 | 0,999 | 0,999 |

| | | PREDICTED CONDITION | |
|---|---|---|---|
| | | Predicted Positive (PP) | Predicted Negative (PN) |
| | | 200 | 99.800 |
| ACTUAL CONDITION | Actual Positive (P) | True Positive (TP) | False Negative (FN) |
| | 100 | 100 | 0 |
| | Actual Negative (N) | False Positive (FP) | True Negative (TN) |
| | 99.900 | 100 | 99.800 |
| | | | |

- Even at the insanely good TPR=TNR=0,999 half of the flagged pictures are false positives
- This practically means that you have to manually screen twice as much pictures than necessary

- Apple(TM) recently proposed(*) an extension to its iOS operating system to test and flag CSAM(**) images on smartphones
- An average smartphone stores no less than 10.000 pictures and there are more than 100.000.000 smartphones in Europe (not all Apple devices)
- Let's assume the extension will operate on 10.000.000 x 10.000 pictures
- The prevalence of CSAM on smartphones is unknown, but should be less than 1/10.000

Well, now ?

(*) Actually, Apple proposal was much more articulated, this is a gross simplification.

(**) Child Sexual Abuse Material

| Total Population (P+N) | Prevalence P/(P+N) | True Positive Rate TPR=TP/P | True Negative Rate (Specificity) TNR = TN/N |
|---|---|---|---|
| 1.000.000.000.000 | 0,000100 | 0,999 | 0,999 |

| | | PREDICTED CONDITION | |
|---|---|---|---|
| | | Predicted Positive (PP) | Predicted Negative (PN) |
| | | 1.099.800.000 | 998.900.200.000 |
| ACTUAL CONDITION | Actual Positive (P) | True Positive (TP) | False Negative (FN) |
| | 100.000.000 | 99.900.000 | 100.000 |
| | Actual Negative (N) | False Positive (FP) | True Negative (TN) |
| | 999.900.000.000 | 999.900.000 | 998.900.100.000 |
| | | | |

- Note that more than 90% of the flagged pictures are false positives
- This actually means that the screening has no practical value
- **AI** cannot overcome the inherent thoughness of screening rare events
- NOTE: the analysis of this case is much harder with the real Apple proposal, but the results are not that different

| Total Population (P+N) | Prevalence P/(P+N) | True Positive Rate TPR=TP/P | True Negative Rate (Specificity) TNR = TN/N |
|---|---|---|---|
| 10.000 | 0,500000 | 0,950 | 0,950 |

| | | PREDICTED CONDITION | |
|---|---|---|---|
| | | Predicted Positive (PP) | Predicted Negative (PN) |
| | | 5.000 | 5.000 |
| ACTUAL CONDITION | Actual Positive (P) | True Positive (TP) | False Negative (FN) |
| | 5.000 | 4.750 | 250 |
| | Actual Negative (N) | False Positive (FP) | True Negative (TN) |
| | 5.000 | 250 | 4.750 |
| | | | |

- Even in a more balanced scenario (e.g. assessing loan risk) there is no guarantee that false negatives are evenly distributed in the overall population
- It is perfectly possible that false negatives are biased against some factor hidden in the training data

## Conclusions

- Don't ask ~~AI~~ ✗ technology to steer the trolley for you, that's only a way to avoid a moral decision
- Ask ~~AI~~ ✗ technology (and technologists) to be clear and transparent on what it can and it cannot do, and to be clear and transparent on figures (precision, recall, specificity...)

And take error into account!

UNIVERSITÀ DI PAVIA

Department of Electrical,
Computer
and Biomedical Engineering

# Steering AI: Legal Challenges and Ethical Standards from an Engineering Perspective

Thursday 18 April
2024

Museo della Tecnica
Elettrica
Via Ferrata, 6
27100 Pavia

*With the patronage*

IL COLLEGIO
FONDAZIONE GHISLIERI

*With the contribution of:*

ReD OPEN
responsibility & design in open ecosystems

*Get the program here*