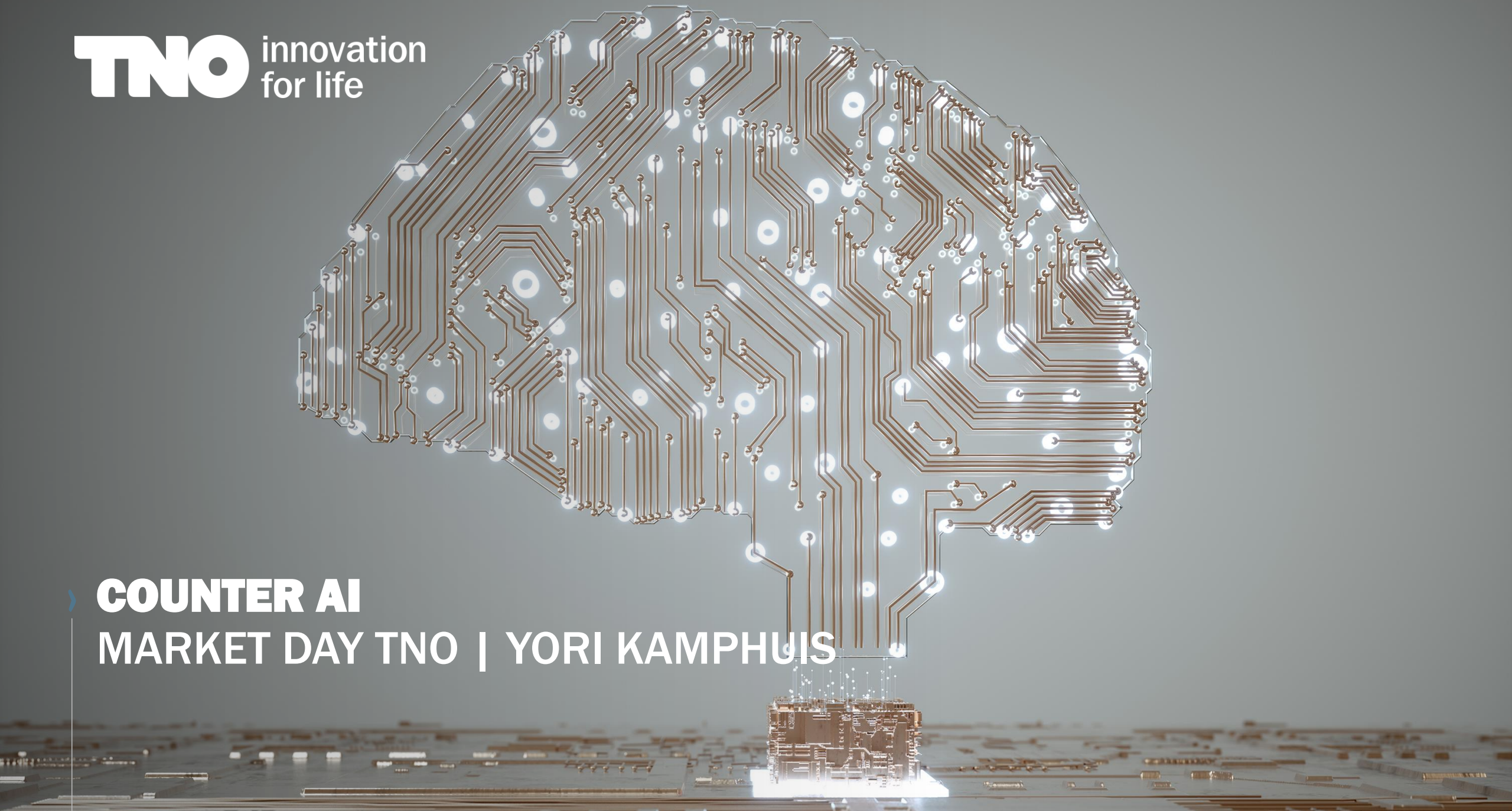
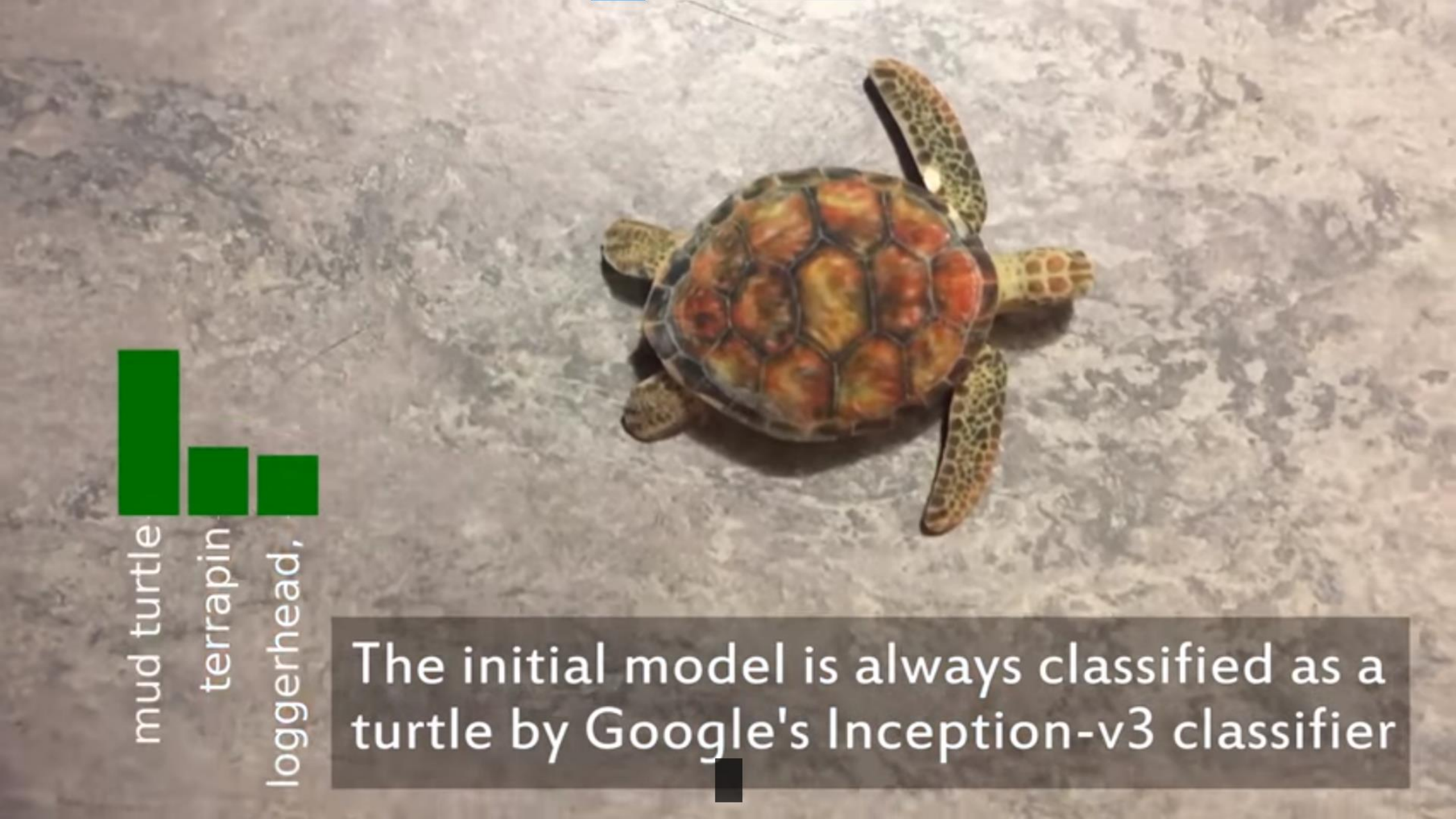
TNO innovation for life



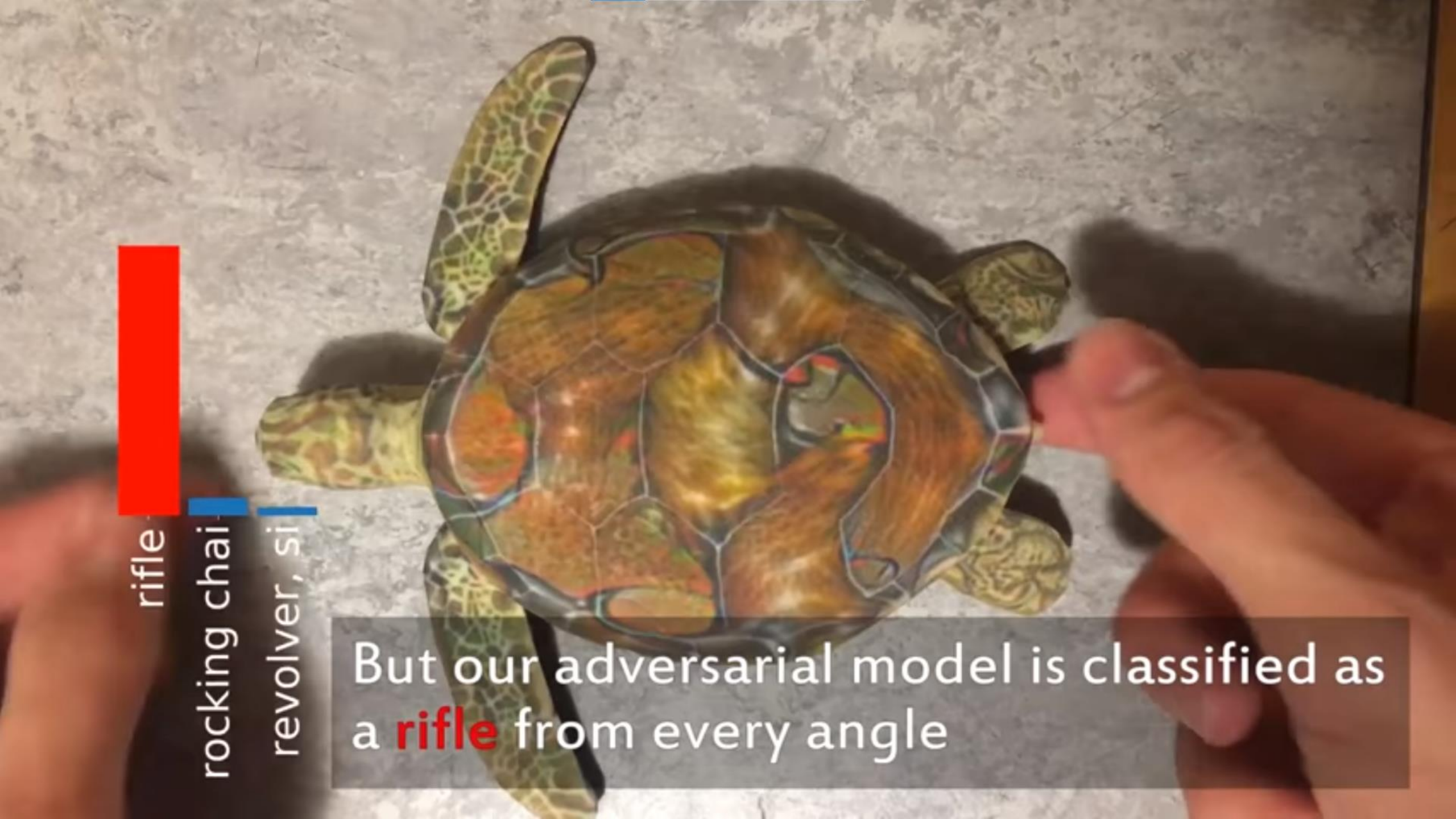
COUNTER AI

MARKET DAY TNO | YORI KAMPHUIS

ARMS RACE



The initial model is always classified as a turtle by Google's Inception-v3 classifier


rifle
rocking chai
revolver, si
But our adversarial model is classified as a rifle from every angle

COMPUTER VISION VS CYBER DOMAIN

IS THIS MALWARE-RELATED, YES / NO?

CORRECT CLASSIFICATION MALWARE: > 95%

CORRECT CLASSIFICATION AFTER ATTACK: < 25%
(I.E. > 75% FALSE NEGATIVES)

ATTACK TYPES

- Poisoning
- Input attack & evasion
- Reverse engineering
- Inference
- Backdoor

UNDETECTABLE BACKDOORS