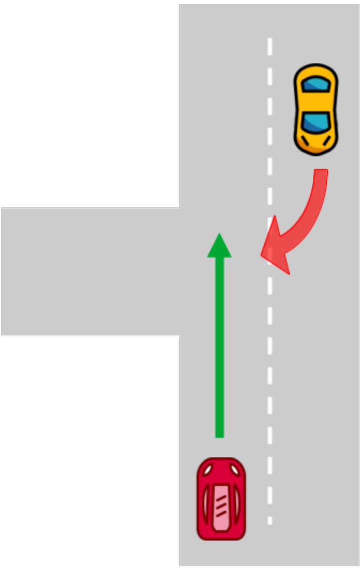
ELEC-E7852 Computational Interaction and Design 2024
Antti Oulasvirta, Danqing Shi
Aalto University

Assignment 4c: Computational rationality (optional, 5p)

This is a bonus task that builds on the Go/No-go task introduced in the Friday lecture.



icons: flaticon

Problem: When we executed the driver model developed in the notebook, we observed that, while the behavior does follow our initial hypotheses, it is not very human-like. The most noticeable discrepancy is a bias toward excess risk-taking. How to design a reward function that produces more human-like behavior?

Task: Reward engineering. Your task is to design a reward function that is more human-like and demonstrate that it works. Start by fixing the risk-taking bias. The present reward function is symmetric: collision is as bad (negatively rewarding) as completing the task (positive reward). You should fix this and then make the model increasingly realistic by redesigning the reward function according to your OWN hypotheses.

Report:

- Reward design (1 page) with reward function, code, and rationale.
- Results from models trained with the new reward functions (1 page with result images)
- Discussion (1~2 paragraphs)

Grading

- Reward function fixes the risk-taking bias +1
- Another un-human-like behavior identified and fixed +1
- Results well-reported with predictions compared against the baseline model +2
- Meaningful discussion +1