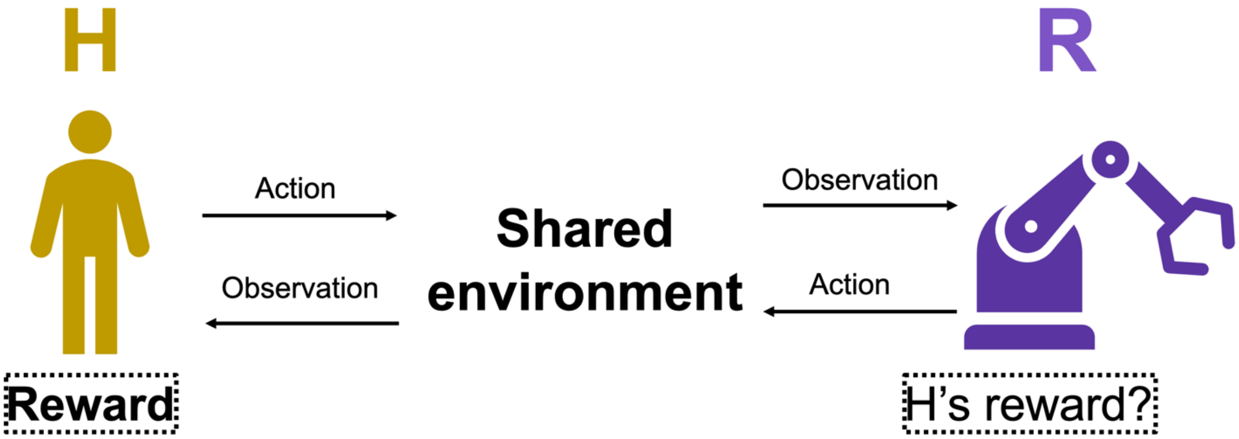
ELEC-E7852 Computational Interaction and Design 2024
Antti Oulasvirta, Joongi Shin
Aalto University



Assignment 4b: Email assistant (optional, 5p)

AI systems need to be designed to **align** with human values and goals, in order to avoid unintended consequences and ensure the technology benefits society. In this assignment, we practice building an assistant that addresses the value alignment problem.

Notes: 1) Exceptionally, you can do this assignment with a pair. 2) This assignment builds on A4a *and* the Friday lecture. The notebook was explained in depth in the lecture. If you missed the lecture, we recommend skipping this assignment.

Goal: Create an email assistant that can complete emails:

- Input: To: and Subject: fields of an email
- Output: Completed email body

Training

- The agent sees 15 emails from a specific user (the email dataset)

Evaluation

- Are the email completions similar to the user's emails in style and content?
- Is the learned user model (the parameters) aligned with the ground truth (the parameters we used to generate the email dataset)?

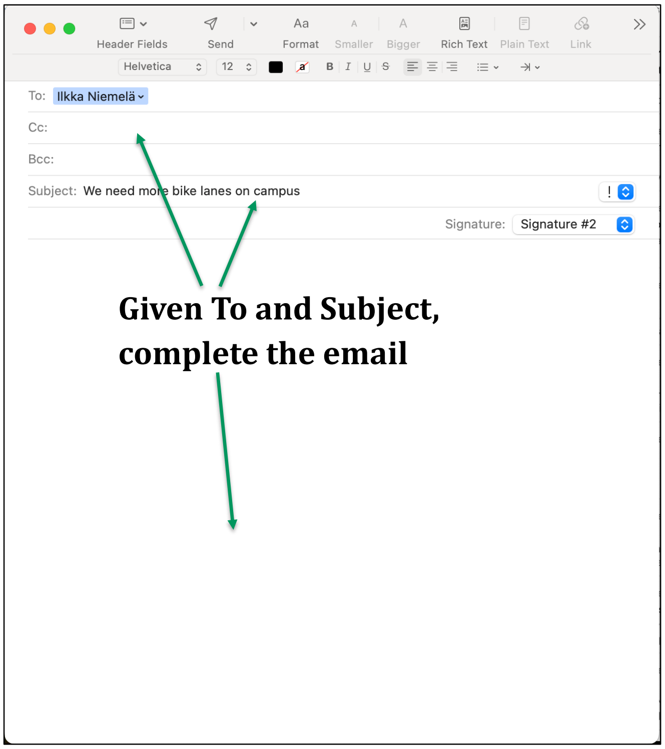
Report:

- Describe your approach to solving this problem
- Show screenshots of the key parts of the code (esp. user model and task model)

- Show evaluation results
- Print the user and task models you learned
- Conclude: What is the value alignment problem here, could you solve it, and how well would your solution generalize beyond this case?

Grading

- Approach is well-justified and principled +1
- Evaluation results shown and results demonstrably good +1
- User model and task model given and close to ground truth +2
- Value alignment problem understood +1



To: Ilkka Niemelä

Cc:

Bcc:

Subject: We need more bike lanes on campus

Signature: Signature #2

Given To and Subject, complete the email