Requirements for Value Alignment Dialogues in Support Agents

PEI-YU CHEN, DR. BIRNA VAN RIEMSDIJK, DR. MYRTHE TIELMAN, DR. CATHOLIJN JONKER, DR. DIRK HEYLEN

TU Delft
UNIVERSITY OF TWENTE.

OVERALL GOAL: FLEXIBLE SUPPORT

- The goal of this project is to realize a behaviour support agent that can provide support in a flexible way such that the support is constantly in line with what the user needs.
- This way, the users do not have to adapt themselves to the technology and can therefore maintain their own space and agency.
- Responsible HI: user and agent work together, shaping their lives in accordance to what they find important.
- Scope and use case and: daily healthy behaviour or healthy lifestyle

CURRENT RESEARCH - A USER STUDY ON ALIGNMENT DIALOGUE

What is misalignment?

- Working definition: "when the support the agent gives doesn't correspond with what the user wants or needs."
- This has two aspects:
  - Because the user model is incorrect
  - Or because the system picks the wrong support

*We are also interested in cases where the right support is given for the wrong reasons.

Notion of Dialogue for Alignment

- When misalignment happens, one effective way to tackle it is to have a conversation between the user and the agent.
- We refer to this as "Dialogue for Alignment:" a dialogue where they talk about the situation and how to solve it for better interaction in the future.

*Note: we use the term "alignment" differently than how it's used in the dialogue system community, in which it's usually referred to the tendency of speakers to reuse aspects of the language they encounter, such as reusing each other's words, the syntactic structure of each other's utterances, or mimicking the other's speech rate, etc.

RESEARCH QUESTIONS

- Which different types of misalignment can happen in a behaviour support agent between the agent and the user?
- How would the end-users want to talk about the misalignment with the agent when it happens?
  - How do different types of misalignment affect how they want to interact?
- What are the effects of different types of the dialogues on the users in terms of their feelings and relationship with the agent?

MISALIGNMENT DIMENSIONS

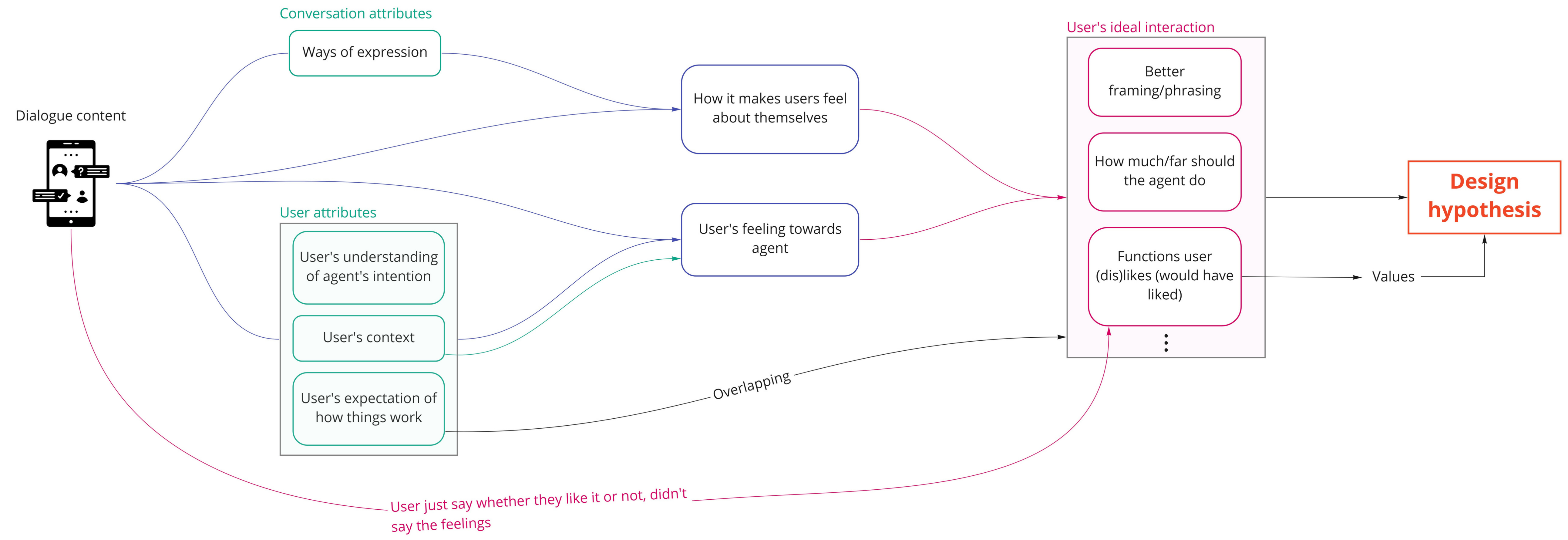
- Scenario setting
  - Prohibition
  - Advice to do something
  - Reminder for not forgetting

  } with different severity
- Triggering condition (what triggers the misalignment)
  - User model is wrong in the first place
  - Context changes
  - User's internal state changes
  - User's desired behaviour changes
  - Right support for the wrong reason

METHODOLOGY

- Two focus groups (6-7 participants)
- Part 1: general questions regarding support agents
- Part 2: six misalignment scenarios and dialogues
  - Different variants of dialogues solving a scenario
  - Questions regarding::
    - Which (part of) variant of the dialogues do you like? Why?
    - The relationship with the support agent: do you feel supported, do they feel in control, etc

RESULTS - CODES RELATIONSHIPS


Conversation attributes
Ways of expression
Dialogue content
How it makes users feel about themselves
User attributes
User's understanding of agent's intention
User's context
User's expectation of how things work
User's feeling towards agent
User's ideal interaction
Better framing/phrasing
How much/far should the agent do
Functions user (dis)likes (would have liked)
⋮
Design hypothesis
Values
Overlapping
User just say whether they like it or not, didn't say the feelings

FUTURE WORK

- Continue to finalize and expand the codes and relationships
- Distill design hypotheses for Alignment Dialogue from this study
- Prototype a conversational agent based on our design hypotheses
- This could then form the basis for a quantitative user experiment to test the design hypothese

REFERENCE

- Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and machines, 30(3), 411-437.
- Ivanova, I., Horton, W. S., Swets, B., Kleinman, D., & Ferreira, V. S. (2020). Structural alignment in dialogue and monologue (and what attention may have to do with it). Journal of Memory and Language, 110, 104052.
- Jonker, C. M., Van Riemsdijk, M. B., & Vermeulen, B. (2010, August). Shared mental models. In International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems (pp. 132-151). Springer, Berlin, Heidelberg.
- Kayal, A., Brinkman, W. P., Gouman, R., Neerincx, M. A., & Van Riemsdijk, M. B. (2013, December). A value-centric model to ground norms and requirements for epartners of children. In International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems (pp. 329-345). Springer, Cham.
- Rabiee, F. (2004). Focus-group interview and data analysis. Proceedings of the nutrition society, 63(4), 655-660.