Computational Persuasion and Explainable AI

Anthony Hunter[1]

Department of Computer Science,
University College London, UK.

July 8, 2021

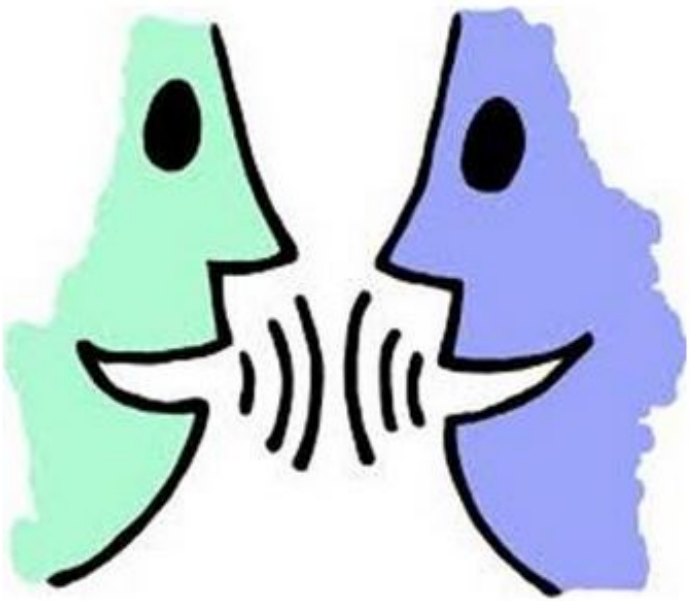
[1]Research done in collaboration with Lisa Chalaguine, Emmanuel Hadoux, Sylwia Polberg, Nico Potyka, and Matthias Thimm

What is persuasion?

Some kinds of interaction surrounding persuasion

- Persuader collecting information, preferences, etc from the persuadee
- Persuader providing information, offers, etc to the persuadee
- Persuader winning favour (e.g. by flattering the persuadee, by making small talk, by being humorous, etc)
- But arguments (and counterarguments) are the essential structures for presenting the claims (and counter claims) in persuasion

What is computational persuasion?



What?

A **computational persuasion system** is an automated system that can engage in a dialogue with a persuadee in order to persuade them to believe something via argumentation.

Why?

The **primary aim** is to persuade via providing useful information and overturning misconceptions.

How?

The system uses a **model of the user** to make a **strategic choice** of moves in an argumentation dialogue.

Computational persuasion in behaviour change

All hospital staff should be encouraged to take the annual flu vaccine.

Hospital staff can infect patients.

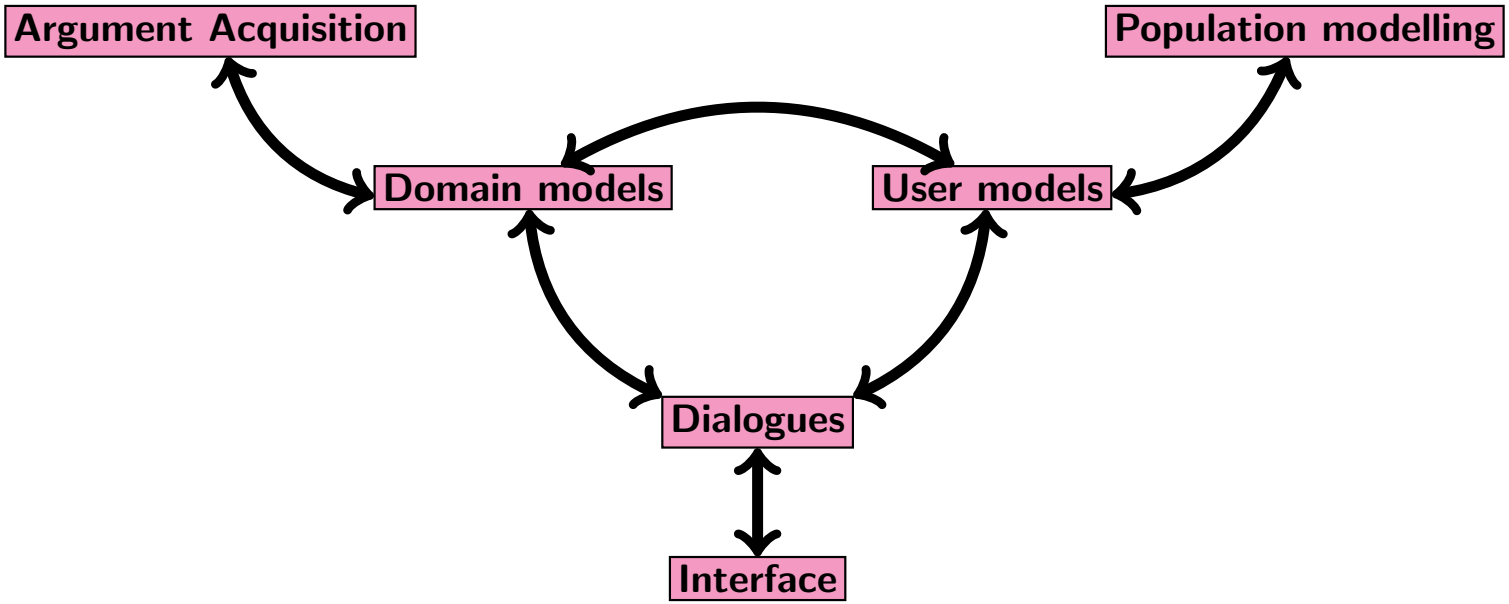
The majority of hospital staff have no face-to-face contact with patients.

Hospital staff are in the same building, breathing the same air, sharing the facilities, touching the same surfaces.

There are disinfectant dispensers that can inhibit the spread of infection.

Infection can also spread with coughing and sneezing.

Overview of computational persuasion project



We explored a range of options in the project

- **Domain modelling:** Abstract argumentation / Probabilistic argumentation
  - **Argument acquisition:** Authoring / Crowdsourcing
- **User modelling:** Beliefs / Concerns
  - **Population modelling:** Beta distributions / Machine learning for predicting beliefs and concerns
- **Dialogue strategy:** Local / Global
- **Interface:** Menu / Free-text

Example of a computational persuasion system

A local strategy for taking concerns into account

- The following is a user argument:
  - Building cycle lanes is too expensive for the city.
- The following are potential counterarguments with concern assignments.
  1. **CityEconomy** Evidence shows that infrastructures for cyclists favour the local economy generating more taxes for the city to use..
  2. **PersonalEconomy** Cycling is cheaper for the citizens than driving or public transportation.
- This ranking over concerns implies counterargument 2 is the best move.

PersonalEconomy > Time > Comfort > Health > CityEconomy

Example of a computational persuasion system

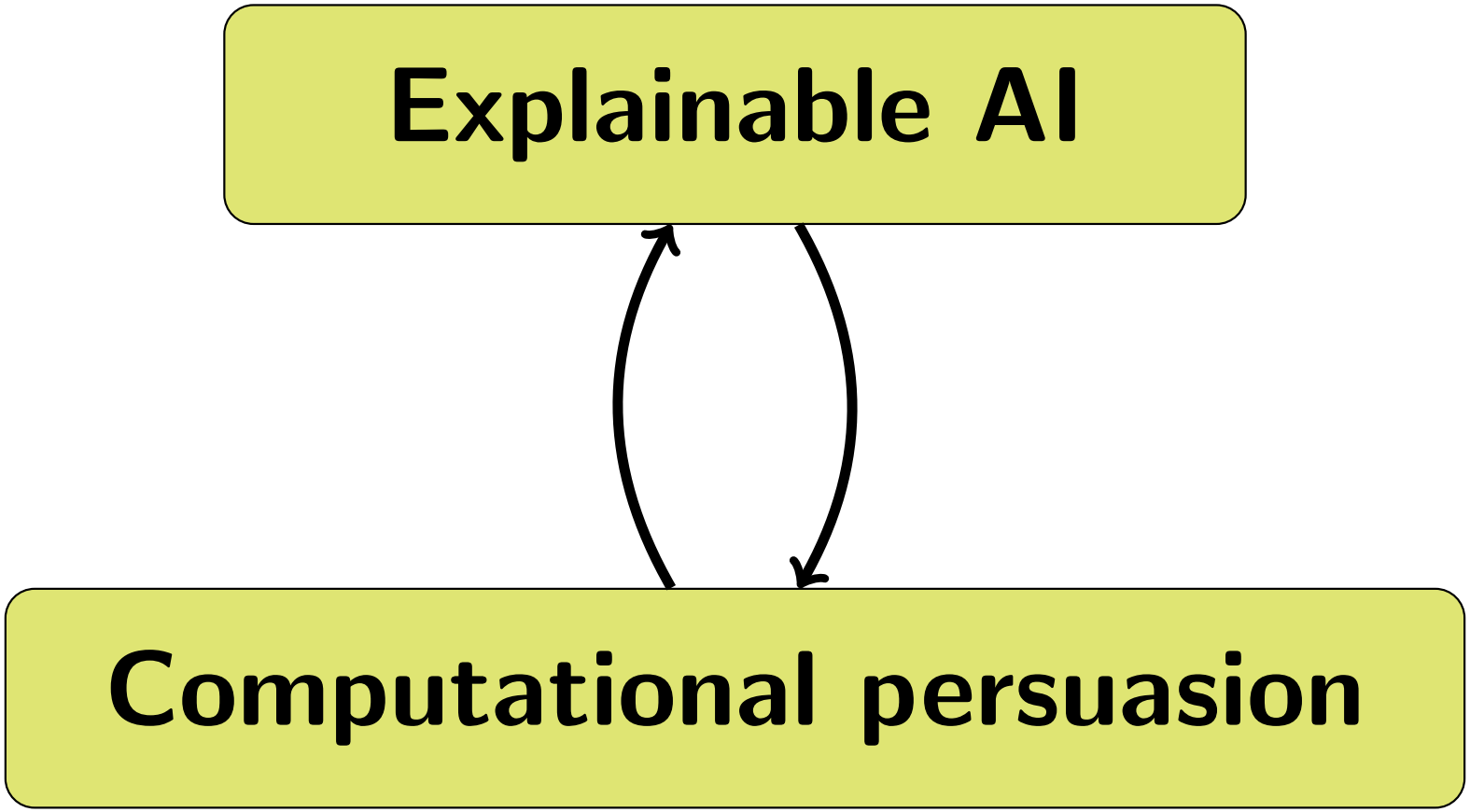
Cycle commuting study

- We used an argument graph with 51 arguments on the topic of commuting by bicycle in the city and 8 concerns (Health, Fitness, Comfort, Time, Personal Economy, City Economy, Environment, and Safety).
- We undertook pre-studies to validate key assumptions
  - Participants tend to agree on assignment of concerns to arguments.
  - Participants give meaningful preferences over types of concern.
  - Participants play by their preferences over concerns.
- We ran a chatbot on the web with 100 crowdsourced participants to persuade them to commute by cycle.

	Strategic system	Baseline system
Negative to positive	6%	6%
Positive to negative	0%	4%

- We obtained a statistically significant improvement in persuasion when compared with baseline system.

Relationships with explanation?



Relationships with explanation?

Example

- Explanation but **not** persuasion
  - The Pfizer vaccine is stored at a very low temperature because it is biologically unstable.
- Persuasion but **not** explanation
  - Take the covid vaccine. Everyone else is.
- Explanation and persuasion
  - Take the covid vaccine. It will substantially lower your risk of getting seriously ill (or worse).

Relationships with explanation?

Does computational persuasion involve manipulation?

- Some non-manipulative approaches
  - Provision of information to persuadee via arguments
  - Overturning misconceptions in persuadee via counterarguments
- Some mildly manipulative approaches
  - Use of emotional arguments
  - Invocation of emotion through choice of language
  - Framing of arguments
  - Use of techniques such as appeal to authority
- Some substantially manipulative approaches
  - Use of dishonesty (e.g. false or unsound arguments)
  - Exploiting biases of persuadee
  - Exploiting ignorance & errors in persuadee
  - Manipulating trust of persuadee in persuader
  - Bullying of persuadee (e.g. gaslighting)

Relationships with explanation?

Explanation for persuasion

- Non-manipulative persuasion is founded on good explanations

Persuasion for explanation

- Maintain a model of the user
- Choose the explanations that are most likely to resonate with the user
- Don't bombard with exhaustive coverage of all arguments
- Be selective in the content that is presented
- Choose the most effective language to communicate with the user

Explainable AI should harness computational persuasion

Conclusions

Conclusions

- **Computational persuasion** based on computational models of argument is a promising approach to technology for behavioural change applications.
- **XAI** could potentially be improved by harnessing ideas from computational persuasion.

More information

www.computationalpersuasion.com