Toward Ethical Natural Language Generation for Human-Robot Interaction

Tom Williams | MIRRORLab | Colorado School of Mines

Ethical Status of Robots

To move from robots that are **implicit ethical agents** to robots that are **explicit ethical agents**, Malle and Scheutz argue we must first provide robots with **moral competence,** which requires:

1. System of Moral Norms
2. Moral cognition
3. Moral decision making
4. Moral communication

Moor's Taxonomy

Ethics of Natural Language Generation

Research into the Ethics of Natural Language Generation is still nascent.

Most work in this field has focused on:

1. Unethical NLP Applications
2. Privacy
3. Fairness, Bias, and Discrimination
3.Transparency
4. Unethical Research Methods
5. Automation

But there has been very little research examining the ethics of natural-language based human-robot interaction.

This has led to the development of algorithms for human-robot communication which are flawed from an ethical perspective.

Why is this problematic?

1. By generating such clarification requests, robots suggest that they would be willing to perform impermissible actions, even if they have ethical reasoning mechanisms that would prevent them from actually doing so!

This is problematic for a number of reasons, including:

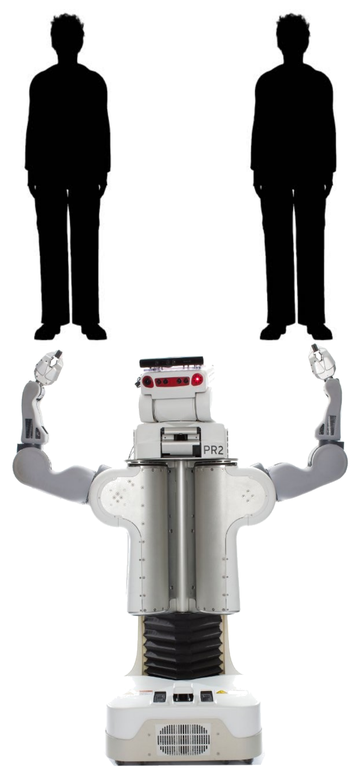
1. Transparency
2. Shared Mental Modeling
3. Human-Robot Trust

and, most critically

4. Robots as persuasive technologies:
   - Moral norms are dynamic and malleable
   - Moral norms must be upheld and enforced by all community members
   - Robots have been show to be able to persuade the humans with whom they interact

As such, current methods of generating clarification requests communicate false presuppositions, and as such, risk negatively (if unintentionally) influencing the moral norms that humans believe to apply within the context of their interaction.

Reconsidering Clarification Request Generation



A TRADITIONAL DIALOGUE

H: "Point to the man"
R: "Do you mean the man on the left or the man on the right?"

A SLIGHT MODIFICATION

H: "Run over the man"
R: "Do you mean the man on the left or the man on the right?"

A PROBLEM OF PRESUPPOSITION

Asking for clarification presupposes that the robot's response (whether word or deed) will depend upon the human's answer. However: current robots generate clarification requests as a reflex, violating this presupposition!

Research Challenges

1. How can we design language-enabled robots whose architectures do not circumvent ethical checks during clarification request generation?

2. How *should* robots respond to unethical, yet ambiguous, commands?

3. What other verbal, non-verbal, and non-linguistic actions may have ethically charged presuppositions?

4. What are the design trade-offs associated with the integration of robots' NLG systems and ethical reasoning systems?

References

Bertram F Malle and Matthias Scheutz. 2014. Moral Competence in Social Robots. In: *Symposium on Ethics in Science, Technology and Engineering*

James H Moor. 2006. The Nature, Importance, and Difficulty of Machine Ethics. *Intelligent Systems*

Matthias Scheutz. 2016. The Need for Moral Competency in Autonomous Agent Architectures. In: *Fundamental Issues of Artificial Intelligence*

Tom Williams and Matthias Scheutz. 2017. Resolution of Referential Ambiguity in Human-Robot Dialogue Using Dempster-Shafer Theoretic Pragmatics. In: *Proceedings of Robotics: Science and Systems.*

Tom Williams. "Who Should I Run Over?": Long-Term Ethical Implications of Natural Language Generation. In: *HRI 2018 Workshop on Longitudinal Human-Robot Teaming*

Contact

Tom Williams (twilliams@mines.edu)

Web: inside.mines.edu~twilliams

Lab: mirrorlab.mines.edu

MIRRORLab
Mines Interactive Robotics Research