The Limits of Moral Turing Tests

Christoph Merdes

Center for Philosophy of Science and Key Qualifications (FAU)

30. June 2020
Agenda

1 Moral Turing Tests

2 Criticism

3 The Way Forward
Agenda

1. Moral Turing Tests
2. Criticism
3. The Way Forward
Morality, just like general intelligence, is difficult to fully specify in explicit terms. Therefore, a test similar to the Turing test has been suggested:

**Moral Turing Test**

- An artificial agent and a human are both presentend a set of moral problems and offer their moral judgment.
- A human interrogator observes these judgments (or potentially chooses the sequence of problems presented) and tries to figure out which agent is human and which artificial by their moral judgments.
- If the interrogator cannot distinguish them above chance level, the machine is considered morally competent.
**Existing Variations**

- **Justificatory MTT:** The interrogator can also engage in a dialogue about the reasoning leading to the judgment.

- **comparative MTT:** The interrogator judges not which agent is human, but which is morally superior, either in
  1. actual behavior in various scenarios
  2. hypothetical scenarios to be judged
Moral Turing Tests

Criticism

The Way Forward
Deception

Pretending Badness

Arnold and Scheutz (2016) point out that the basic MTT may incentivize the machine to deceive the interrogator by being less morally competent.

- Conflict between deceit and moral competency (BUT: Consequentialist ethics)
- Suppression of superior artificial morality (BUT: Speculative, and questionable whether rationally inaccessible morality can be binding.)
- This Problem has been suggested to be addressed by the cMTT.
A Superior Option?

Verification
Arnold and Scheutz (2016) suggest verification, that is, formally proving the compliance of a system with an explicit, formalized moral theory.

- Strictly speaking impossible for a machine operating in an open environment.
- Does not take into account the variety of competing theories.
- Ignores the fact that various views exist that deny the possibility of an explicit, formal statement of what morality entails (e.g. particularism (Dancy, 2017)).
Content

Observation: Widespread, deep-running moral disagreement

- MTT reproduces interrogator’s values
- MTT may fail due subject/interrogator disagreement
What is Moral Competency?

Content

Observation: Widespread, deep-running moral disagreement
- MTT reproduces interrogator’s values
- MTT may fail due subject/interrogator disagreement

Form

Wallach and Allen (2008): jMTT solves disagreement, because it gives us the reasoning
- Nobody’s values
- Reduces morality to intelligence (self-defeat)
What is Moral Competency?

Content

Observation: Widespread, deep-running moral disagreement

- MTT reproduces interrogator’s values
- MTT may fail due to subject/interrogator disagreement

Hypothesis: This puzzlement is explained by the incoherence of the concept of moral competency assumed. Every MTT could be understood to define a coherent interpretation, but there is no clear concept they could even try to capture correctly.

Form

Wallach and Allen (2008): jMTT solves disagreement, because it gives us the reasoning

- Nobody’s values
- Reduces morality to intelligence (self-defeat)
Custom Morality

The artificial moral agent could, to sidestep disagreement, be morally calibrated to its master (who could play the interrogator role in a cMTT).
The artificial moral agent could, to sidestep disagreement, be morally calibrated to its master (who could play the interrogator role in a cMTT).

- Very few cases where acceptable (cf. autonomous vehicles)
- Not actually moral at all
- Highlights the fact that we implicitly assume MTTs to be equipped with some veil of ignorance (Rawls, 1958)
An MTT can also embrace disagreement:

**social MTT:** Instead of a single interrogator, a group of interrogators is employed, and their judgments aggregated in some way into a compromise.
An MTT can also embrace disagreement:

**social MTT:** Instead of a single interrogator, a group of interrogators is employed, and their judgments aggregated in some way into a compromise.

- How to aggregate individual judgments?
  - **Deliberation:** May fail, psychological and power dynamics.
  - **Immediate Aggregation:** Prone to impossibility results (List et al., 2011).

- Who ought to be in that group in the first place?


