Can we forgive a robot?

Michael Nagenborg

Can we punish a robot?

- Peter Asaro (2012) points out in "A Body to Kick, but Still No Soul to Damn" that robots have no sensations, thus do not experience harm and cannot be punished.

- In contrast: Daniell Dennett (1998) argued, that robots might indeed have something like artificial sensations.

Overview

- Why forgiving matters in human-human relations
- Why forgiving matters in human-robot relations
  - *Forgiveability* as a condition for ascribing responsibility
  - Forgiveness as a litmus test of the human-robot relations
  - Forgiveness as a (human) virtue for living with robots
- Summary and outlook

Part 1

Why forgiving matters in human-human relations

(inspired by Hannah Arendt)

Vita Activa

- *Making promises* and *forgiving* are central activities in human existence.

- The ability to make promises and accept them is a central mechanism for dealing with the unpredictability of human beings. Through these acts we create "islands in a sea of uncertainty" (Arendt 2002: 313).

- „Could we not forgive each other, that is, relieving each other of the consequences of our actions, our ability to act would become limited to a single act whose consequences would literally haunt us until the end of our lives […].“ (Arendt 2002: 302)

Forgiveness and freedom

- Person X promises person Y to do Z.
- Person X breaks his / her promise (and does not do Z).
- What can Y do?
  - Option 1: Negative reaction (‚punishment of Y‘)
  - Option 2: Forgive Y for not doing Z.

Forgiveness and freedom

Two challenges:

1. If X has only one option to react to Y's misdoing, Y's failure to do Z determines the reaction of X.

2. To make matters worse: ‚punishment' is a morally ambivalent act. Without the option to forgive Y, X is bound to *harm* Y.

Part 2

Why forgiving matters in human-robot relations

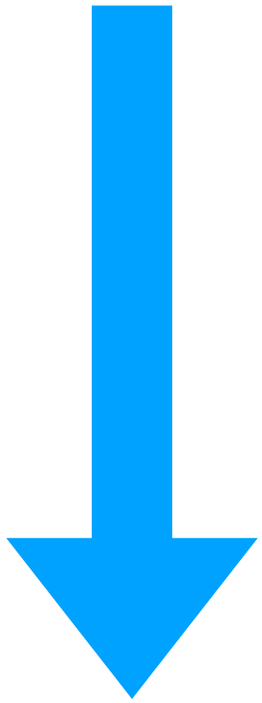
Forgiveability as a condition for ascribing responsibility

Do robots *think*?

Likewise according to this view the only way to know that a man thinks is to be that particular man. It is in fact the solipsist point of view. It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe "A thinks but B does not" whilst B believes "B thinks but A does not." instead of arguing continually over this point it is usual to have the polite convention that everyone thinks.

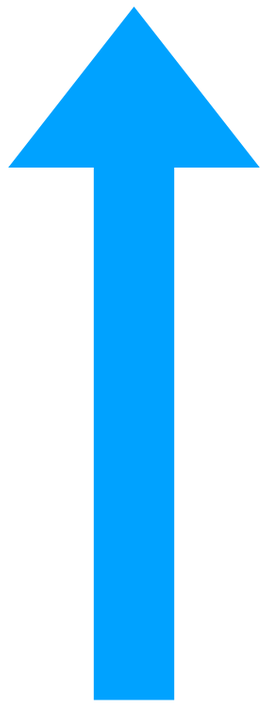
Alan Turing (1950)

Responsibility



Forgiveness

Responsibility



Forgiveability

Forgiveness as a litmus test of the human-robot relations

Two trivial cases

- Robots which are considered to be „mere machine.“

- Robots which have all the relevant properties of a person.

Minimal conditions

- The robot can make *unpredictable* choices.
  - Alternative actions are possible.
    - It’s impossible (for practical reasons) to predict for a human the choice of the robot.
- The robot has *reasons* for its choice.
  - Actions are more than random acts.

Challenges

- Even if a robot has reasons, there is no guarantee that it will be good reasons by human standards.

- Reason for forgiving a human person: „If I would have been you in that situation - I may have done the same.“

  - Human-human forgiveness: The common nature of the two parties can be taken for granted.

  - Human-robot forgiveness: The lack of a common nature or reasoning becomes problematic.

Prospect

- **Litmus test:** If human beings start to sincerely and honestly forgive robots, it would be a strong indication of a fundamental shift in the relationship between human beings and robots.

- Arendt (2002): „I do not forgive you, because of what you did, but because of who you are.“ (My translation)

    - Forgiving a robot: „I do not forgive you, because of what you did, but because of *what* you are.“

Forgiveness as a (human) virtue for living with robots

Sovereignity and foregiveness

- Forgiving is a sovereign act.
    - One may hope for forgiveness, but there is no right to be forgiven.
- Reminder: Arendt
    - Forgiveness as alternative option to punishment.
    - I only remain free, if I am not forced to forgive under specific circumstances.

Sovereignity and foregiveness

- Even if forgiving is a sovereign act, it might be reasonable to be forgiving.

  - (Machiavelli: Instrumental value of forgiveness)

- Moral ambivalence of ‚punishment‘ - without being able to forgive, we are bound to do ‚harm‘ to others.

  - In other words: If we can’t forgive, we condemn ourselves to do ‚harm‘ (even if the robot does not feel pain).

2nd class moral responsibility

- What would happen, if we do not grant a robot full moral agency?

  - The robot can be blamed, but can not be forgiven.

    - Since nothing has changed for human-human forgiveness, we end up with two classes of moral agents: Both kinds of moral agents can be blamed, but only one can be forgiven.

- Therefore, we should not ascribe moral responsibility to an artificial agent unless we are willing to forgive it.

Summary and outlook

Summary

- Forgiveness provides with an interesting theoretical lens to look at human-robot relations:
- If humans start to forgive robots, this should be seen as an indicator for a fundamental shift in human-robot relations.

- We should be very careful about ascribing moral responsibility to a robot, if we are not willing to forgive that robot.

Outlook

- Still unclear to me, why Arendt does consider ‚forgiveness‘ only in the context of politics and not in the context of complex works.

- My research is based on the German discussion on the role of „forgiveness“ in Ethics. Open question:

  - What about different accounts of „forgiveness“ in Western and other traditions?