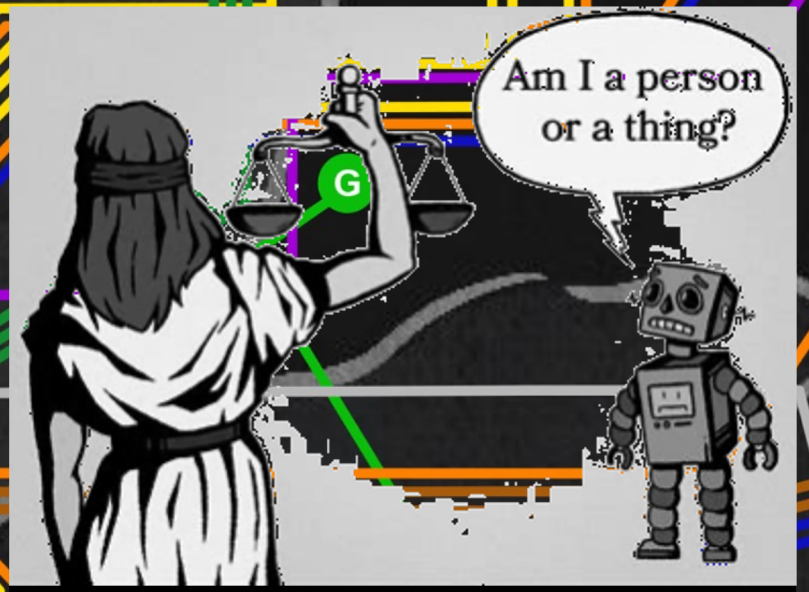
Situating Machines within Normative Practices



Anna Strasser & Mike Wilby

NYU

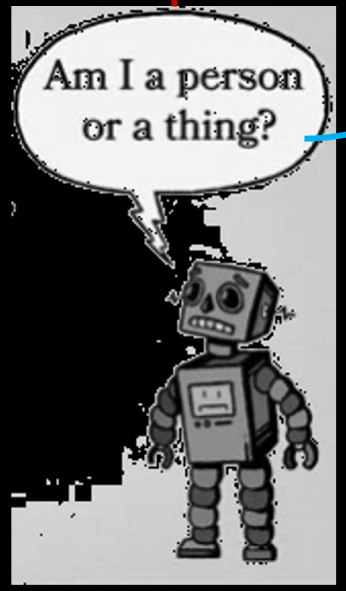
MANHATTAN

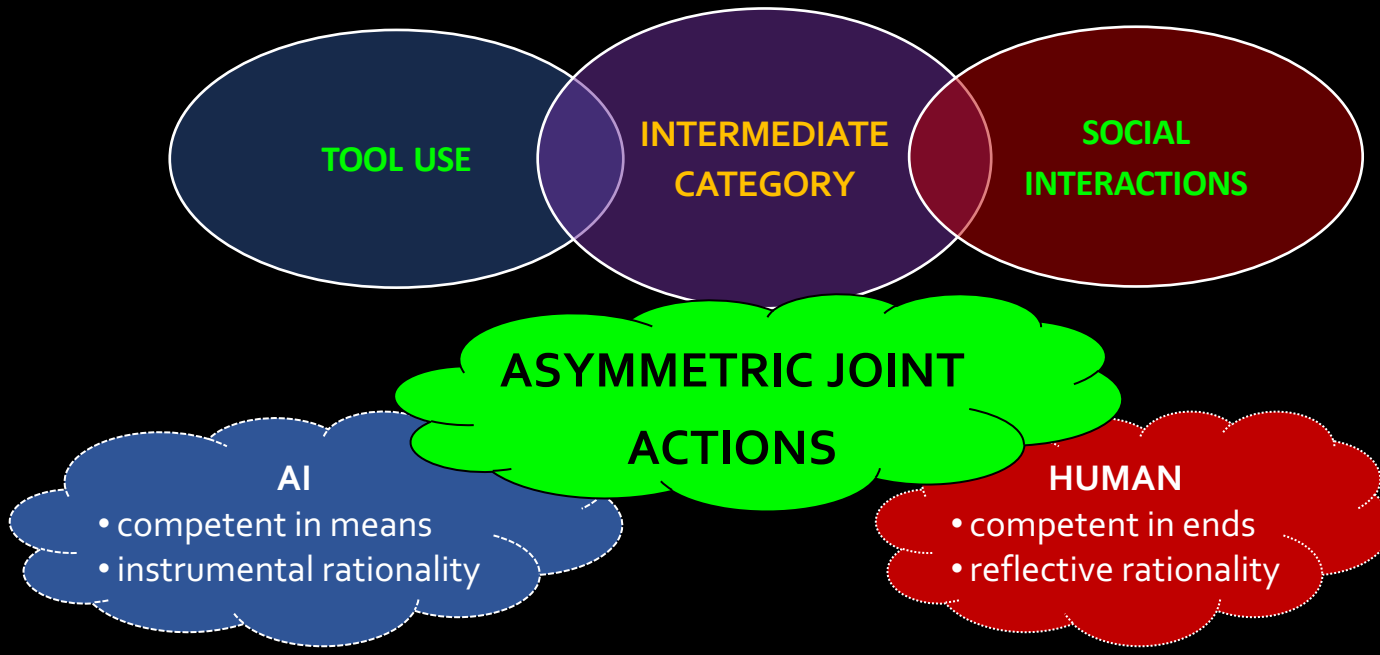
slides can be downloaded at https://www.denkwerkstatt.berlin/ANNA-STRASSER/TALKS

1 Artificial systems as genuine agents

full-blown intentional systems → too inflationary

mere tools → too reductive





The AI-Stance

Treat AI as highly competent in terms of means (instrumental rationality), but virtually no competence in terms of determining ends (reflective rationality).

DENKWERKSTATT BERLIN

New Responsibility Gaps?

HOMING IN ARTIFICIAL SYSTEMS INTO THE REALM OF SOCIAL OR QUASI-SOCIAL INTERACTIONS
IT IS NOT CLEAR HOW TO TREAT QUESTIONS REGARDING RESPONSIBILITY.

IF
- artificial systems act & do not merely behave

THEN
- some of the actions themself might be genuinely morally blameworthy

BUT AT FIRST SIGHT

❖ no viable candidate for moral responsibility
(only a minimal form of agency but no moral agency

New responsibility gaps in human-machine interactions

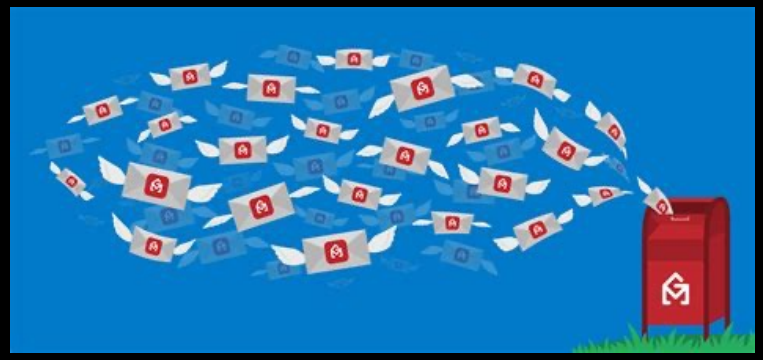
- neither the AI (lacks moral capacity)
- nor the HUMAN (lacks agential responsibility)
- nor the manufacturer (cannot anticipate the actions of the AI)

can be responsible

DENKWERKSTATT BERLIN

A morally significant Human-Machine Interaction

WORKING WITH A MODERN AI SYSTEM ON A MORALLY DELICATE TASK



AI system trained by a human that is able to compose and send 100s of emails at a go using prepared information about the recipients

JOINT TASK

sending emails to inform vulnerable people of either good or bad news that will change their lives unalterably with sharp time constraints

AI:

- composing & sending emails

HUMAN:

- monitor the emails
  - at a rate that is acceptable for the job
  - but also distracted by other things

SOMETHING GOES WRONG

- several of the batches of emails from the AI sending out inappropriate and inaccurate information

→ **going to be highly harmful & hurtful to the individuals**

➢ Significant harms have been incurred.

WHO IS MORALLY RESPONSIBLE FOR THIS?

DENKWERKSTATT BERLIN

The coupled HMI system, insofar as it is interpretable as a full intentional system, can, as a whole, be held morally responsible.

How can we answer the question of how the blame is – or should be – distributed among the actors?

HUMAN SIDE

Prospective & retrospective responsibility

PROSPECTIVE RESPONSIBILITY FOR A PROJECT DOES NOT ALWAYS ENTAIL RETROSPECTIVE RESPONSIBILITY FOR WHEN SOMETHING GOES WRONG WITH THAT PROJECT

- on-the-loop supervisor → prospective responsibility

  example building site
  - If one of the builders goes on a wild, murderous rampage → retrospective responsibility for the murderous outcomes clearly stands with the murderer

- HMI example
  - ➢ prospective responsibility of the on-the-loop supervisor does not necessitate retrospective responsibility
    (as long as the supervisor did what was expected)

But if the supervisor, doesn't have retrospective responsibility, who (or what), if anything, does?

WHO IS MORALLY RESPONSIBLE FOR THIS?

DENKWERKSTATT BERLIN

Minimal Moral Responsibility

AI SYSTEMS CAN BE INVOLVED IN THIS MINIMAL FORM OF MORAL PRACTICE.

Moral Responsibility:
Full-blown moral responsibility requires participation within full-blown practices of moral blame. Such full-blown practices involve complex cognitive, affective, and behavioral responses to the actions and attitudes of others (Strawson 1963).

The AI system is not capable of engaging in such practices.

A more minimal notion of moral blame

- involves just the behavioral component, which Scanlon calls a "modification" of the relationship, involving the "withdrawal of trust" (Scanlon 2015, p. 93).

Applying this to AI systems

- we suggest that a new social practice
  → a normatively appropriate 'withdrawal of trust' presupposing that after each human-machine interaction, there will be a procedure of evaluation and the human is responsible for checking whether the AI did learn from this evaluation.
- And this can serve as a basis of withdrawal of trust.

DENKWERKSTATT BERLIN

Take home message

sometimes you do not need to assume consciousness

If you were a machine, and we had (quasi-)social interactions (asymmetric joint actions)

- I would treat you with care, but I will make you responsible for messing up with your part, concerning which I have no chance to intervene.

- You will have retrospective responsibility.

- I will carry the prospective responsibility.

TO JUSTIFY ASSIGNING RETROSPECTIVE RESPONSIBILITY TO YOU, I SUGGEST A NEW SOCIAL PRACTICE, NAMELY, THAT THERE WILL BE AN EVALUATION AFTER EACH INTERACTION AND IF YOU DO NOT LEARN FROM THIS ...., THEN I WILL NOT INTERACT WITH YOU AGAIN AND WITHDRAW TRUST.

DENKWERKSTATT BERLIN

END