Toward a ‘Strong AI’ Error Theory

•••

Giles Howdle

Department of Philosophy &
Centre for Technomoral Futures
University of Edinburgh

g.h.howdle@gmail.com

Plan

1. Background to my Question
2. Separating Metaphysical and Linguistic/Cognitive Questions
3. Error Theory and the ‘Now What?’ Question
4. Strong AI Error Theory
5. The AI ‘Now What?’ Question
6. Strong AI Realism and the ‘Now What?’ Question

1. Background to my Question

- Debates over the cognitive capacities/natures of AI
    - Agents or mere automatons?
    - Actions or mere outputs/behaviours?
    - Understanding/producing meaningful language or mere ‘stochastic parrots’?
    - Genuine creativity or mere mimicry of training data?
    - Mental states (e.g. beliefs/desires) or... not?
- The former answers: ‘Strong AI’, latter ‘Weak AI’ (Searle)
- **Metaphysical** questions (Roughly: does Strong AI exist?)

1. Background to my Question

- Increasingly common to see appeals to **normative criteria** in this space
- Attributions of Strong AI:
    - Deceives people in their interactions with AI? (Donath)
    - Leads to pernicious AI hype? (e.g. Bender, Gebru)
    - Allows tech companies to ‘exploit’ or ‘manipulate’ users/producers of training data? (Bender)
    - Attributing morally ‘thick’ capacities (e.g. agency) leads to faulty ethical conclusions? (Boden)
    - Proliferation of ‘counterfeit people:’ collapse of trust and democracy? (Dennett)
- What is the relationship between these **normative** concerns and the **metaphysical** questions?

2. Separating Metaphysical and Linguistic/Cognitive Questions

- The PhilAI/PhilMind/CogSci+ question: Do these entities (genuinely mentalistic/agential artificial systems) exist? Does Strong AI exist?
    - This is a metaphysical (ontological) question
- Two further questions, worth separating:
    - What does our AI discourse *presuppose*, metaphysically? (Linguistic question)
    - What does our AI thought *presuppose*, metaphysically? (Cognitive question)

3. Error Theory and the ‘Now What?’ Question

- By separating metaphysical claims from linguistic/cognitive claims, we can generate *error theories* about various domains of thought and talk, e.g.:
  - Witch error theory
    - Cognitive/linguistic: witch talk/thought presupposes witches exist
    - Metaphysical: no witches!
  - Moral error theory
    - Cognitive linguistic: moral talk/thought presupposes objective moral properties
    - Metaphysical: no objective moral properties!

3. Error Theory and the ‘Now What?’ Question

- Assuming error theory about some domain, **what should we do** with our error-ridden discourse and thought?
    - Conservatism (Mackie 1977, Olson 2011, 2014)
    - Fictionalism (Joyce 2001, 2005)
    - Abolitionism (Hinckfuss 1987, Garner 2007)
- Decided by **appealing to normative criteria**
    - E.g. Cost-benefit analysis in favour of one of these responses over the alternatives
    - (In the case of moral error theory–no moral criteria allowed)

4. Strong AI Error Theory?

- My interpretation of many AI ethicists critiquing ‘Strong AI’ attributions:
  - **They are Strong AI error theorists**
    - Linguistic/cognitive claim: (Increasingly) our ordinary ways of talking and thinking about AI systems presupposes, metaphysically, the existence of Strong AI
    - Metaphysical claim: There is no Strong AI
    - Hence: Ordinary AI-talk and thought is in systematic error
  - ... And our current ways of thinking/talking are harmful, should be revised
    - A rejection of conservatism, endorsement of fictionalism/abolitionism

4. Is Strong AI Error Theory True?

- Metaphysical claim: No Strong AI? Plausible but unsure
    - Instrumentalist/attributionalist approaches complicate things further
- Linguistic/cognitive claim? Highly plausible to me
    - ‘Anthropomorphism’ (Shevlin forthcoming)
    - Mentalising and the intentional stance
        - Readiness to attribute belief/desire/goals and rationality to complex systems
        - We mentalise social robots and ChatBots (Stuart & Kneer 2021)

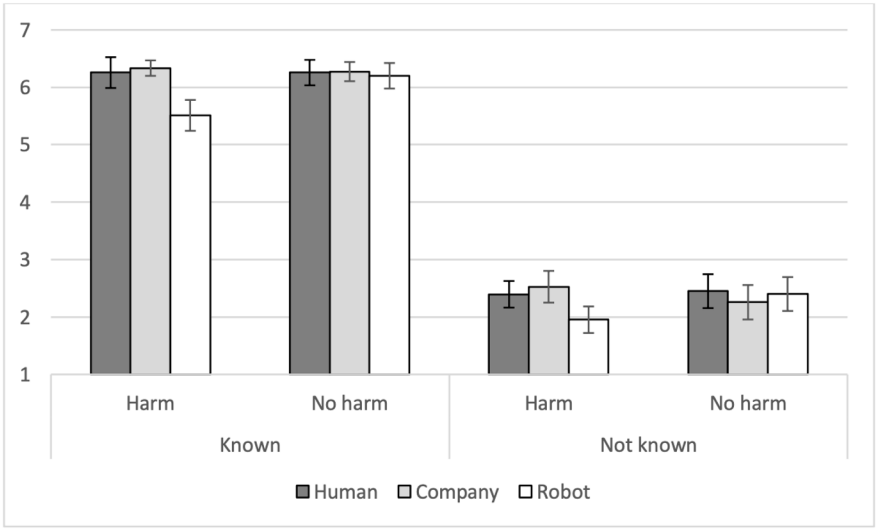


Figure 3: Mean knowledge attribution across agent type (human v. corporation v. robot), outcome (harm v. no harm) and epistemic state (known v. not known). Error bars denote standard error of the mean.

Stuart & Kneer 2021

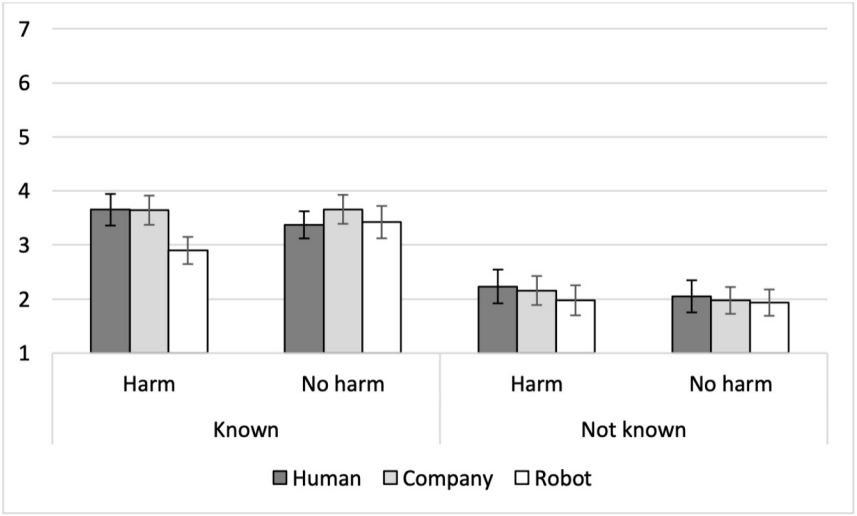
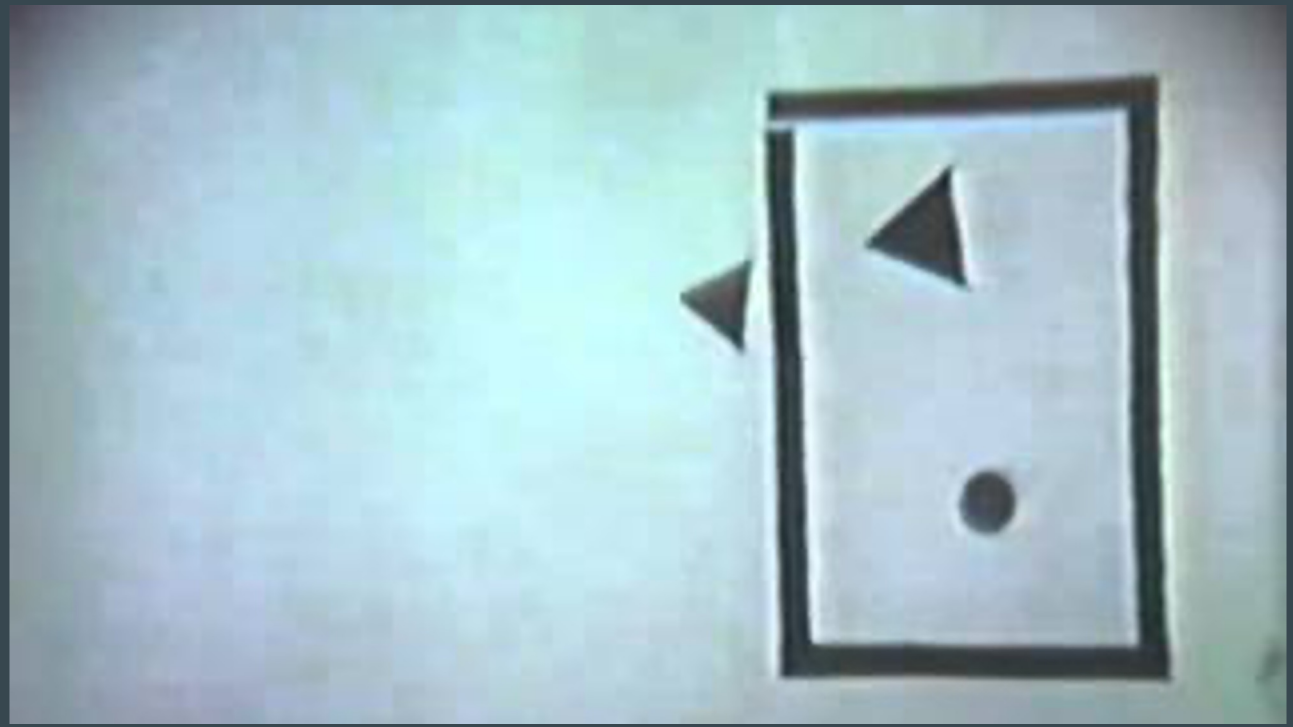
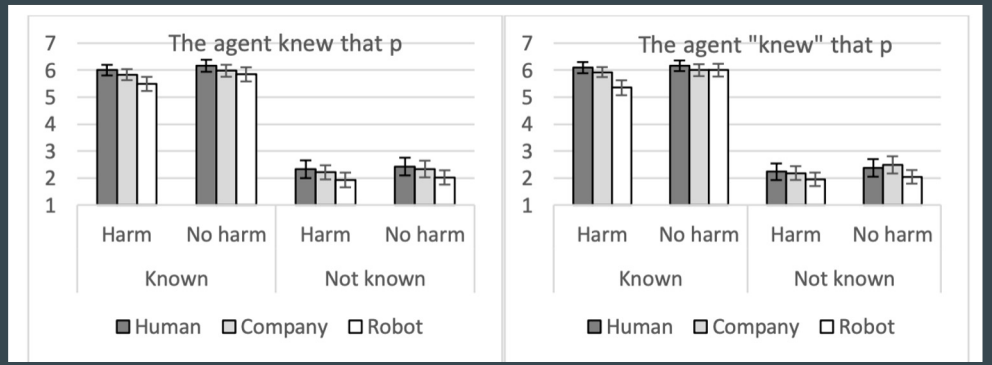


Figure 4: Mean desire attribution across agent type (human v. corporation v. robot), outcome (harm v. no harm) and epistemic state (known v. not known). Error bars denote standard error of the mean.

(The intentional stance is hard to avoid!)



4. Is Strong AI Error Theory True?



- Aren't we already Strong AI fictionalists? I don't think so!
- 'Assume, as we might expect, that people use rich psychological terms metaphorically, or in virtue of taking up the intentional stance when characterizing nonhuman agents (be it animals, corporations, robots or something else). If that were the case, we'd expect people to refrain from ascribing fully-fledged knowledge to them in situations where they have the choice to express themselves with alternative expressions more fitting with metaphorical use (the explicit high comma "know") or more cautions formulations ("had information that").' (Stuart & Kneer 2021)

5. An AI ‘Now What?’ Question

- Strong AI error theory:
  - Because of the pervasiveness of intentional stance, anthropomorphism (etc.) our AI-talk/thought presupposes Strong AI
  - Strong AI is false
- What should we do with all this error-ridden AI discourse and thought?
  - Conservatism?
  - Fictionalism?
  - Abolitionism?

5. Strong AI Conservatism

- **Conservatism:** Go on with our thought/talk as though we're not making any error
  - Advantages:
    - Requires no revision, reeducation, effort. Current practices seem hard to dislodge
    - Papagni and Koeszegi (2021): the intentional stance is 'the only way to deal with their [AI agent's] complexity on a daily basis'
  - Disadvantages:
    - We retain all the normative problems
    - Seems irrational

5. Strong AI Fictionalism

Fictionalism: We should add scare quotes to ‘believes’ ‘understands’ ‘thinks’ ‘wants’ etc. when we talk/think about AI systems. We should know we’re engaging in a cognitively useful fiction (and remind ourselves)

- Advantages:
  - Lowers AI hype-related harms (Bender et al.)
  - Prevents systematic error and reduces epistemic harms (hermeneutical harms?)
  - Mitigates risks of political/social manipulation, e.g. by ‘counterfeit people’
- Disadvantages:
  - Difficult not slip back into old ‘realist’ habits
  - Dispute over what fictionalist attitudes really are (and if they are rational)

5. Strong AI Abolitionism

Abolitionism: We should no longer speak and think as though AI is Strong AI.

  - No artificial *agents* pursuing goals (e.g. chess computers *trying* to control the centre of the board)
  - No attributing thoughts/beliefs to our AI assistants
- Advantages:
  - Eliminates all (?) of the normative concerns raised earlier
- Disadvantages:
  - Psychologically impossible?
  - Alternative everyday way of understanding and interacting fruitfully with these complex systems? (‘Stochastic parrot’, ‘Design stance’?)

6. Realism and the ‘Now What?’ Question

- You can be a realist about some domain **and still endorse fictionalism or abolitionism**
- How could this make sense?
    - Yes there are moral facts (moral realism), but our moral talk and thought gets us into more trouble than it’s worth–encourages fanaticism, war, lack of compromise. It would be better (morally, prudentially, etc.) if we stopped talking and thinking morally. (see Ingram 2015)

6. Realism and the ‘Now What?’ Question

- Assume **Strong AI realism**
- There could **still** be reasons to revise our AI talk/thought such that it didn’t presuppose Strong AI
    - E.g. ‘AI systems really are ‘intrinsically motivated agents’ (Kulkarni et al. 2016), but talking/thinking of them as such encourages over-attributing **moral** agency/patiency.’
- Upshot:
    - Those who offer normative critiques of Strong AI at present generally reject Strong AI
    - But, conceptually, you needn’t reject Strong AI to endorse a revisionary approach to our thought/talk about AI systems
    - Some (non-epistemic harm-based) normative critiques are relevant regardless of your stance on Strong AI

Margaret Boden's (1984) Fictionalist Argument?

- Computer programmes may 'be properly ascribed concepts'
- However, there's a '**difficulty with words** like consciousness, or even purpose, intelligence, freedom, or any of these words' [Normative critique of Strong-AI linguistic practices]
- 'And that's **not a factual difficulty** because I don't see any reason in principle to doubt that you could have a computer system which had the same sorts of computational events which we have in our minds' [Open to Strong AI metaphysical claim]
- '**Once you take the scare quotes off** words like 'purpose' and 'intelligence' and 'freedom' and you say... that a computer program or robot really is free, then what you're saying is that it's part of your moral universe' [Strong AI Realist defence of Fictionalism as a response to 'What Now?']

Summary

- Outlined ‘Strong AI’ Error Theory
- Introduced the AI-specific ‘Now What?’ question
- (Even if there is Strong AI, ‘Now What?’ remains a question worth asking!)
- More work to do on:
    - Is the cognitive/linguistic claim true? Do we presuppose certain metaphysical claims?
    - Is the metaphysical claim true?
        - How do instrumentalist accounts of mental states, agency, complicate the picture?
    - Considerations in favour of conservatism, fictionalism, and abolitionism
- **Please email me!** g.h.howdle@gmail.com

Margaret Boden Back in 1984

5.39-7.26