AI and Illusionism as a Risk to Altruism

Incl. musings on the Fermi Paradox

The Science of Consciousness Conference 2020, Sept 14-20

Short version

Florian Habermacher, PhD

Sept 2020

Update June 2022

Disclaimer

Views are Author’s own, unrelated to affiliations.

Comments welcome at

florian.habermacher@gmail.com

Short version. More detailed version available from the author

Intro

Premises

- P1 Weak Computationalism
- P2 Aware AI

Hypotheses

- H1 AI ⇒ Illusionism
  - a. Reality: AI suggest Illusionism is True?
  - b. Popular: AI makes Illusionism more Popular?
- H2 Illusionism ⇒ Less Altruism?
  - a. Reality: Illusionism *Justifies* Less Altruism?
  - b. Popular: Illusionist Reduces Altruistic?

Implications

- Altruism vs. Cooperation
- Policy
- Consider Trump

Wilder Speculation

- On the Fermi Paradox
- An Altruism Paradox

Definition Illusionism

View that:

Phenomenal experience = Illusion created by our brain

What we **“believe” to feel** = **Computation.** Explained by **basic physics**.
No “intrinsically valuable feeling”.

Background on Illusionism

Illusionist positions by various figures in consciousness studies.

Dennet, Frankish, Kammerer, …

But:

Defies what we most obviously think – or feel – to know.

Illusionism “seems not to take consciousness seriously”

Chalmers 1995, noting – nevertheless – that the theory is to not be dismissed light-heartedly

Here:

No encompassing pro-illusionism argumentation, but **few illustrations as to why illusionism might be(come) more appealing than one might spontaneously imagine**

Overview

Emergence of AI emphasizes closeness of brain and machine

Exposure to AI may popularize Illusionism? Rightly or wrongly

Illusionism can severely curb Altruism? Rightly or wrongly

Implications

Social risk; Policy need

Trump

Wilder Speculation

Fermi Paradox; AGI may kill itself off before it exists; Altruism Paradox

Intro

Premises

P1 Weak Computationalism

P2 Aware AI

Hypotheses

H1 AI ⇒ Illusionism

a. Reality: AI suggest Illusionism is True?

b. Popular: AI makes Illusionism more Popular?

H2 Illusionism ⇒ Less Altruism?

a. Reality: Illusionism *Justifies* Less Altruism?

b. Popular: Illusionist Reduces Altruistic?

Implications

Altruism vs. Cooperation

Policy

Consider Trump

Wilder Speculation

On the Fermi Paradox

An Altruism Paradox

Premise 1: Weak Computationalism

Weak Computationalism:

“Dualism is dead; **Physical neurons etc. directly link my senses to tongue & legs, explain my moves**. But my genuine phenomenological experience convinces me **there’s sth beyond physical computation; it’s the hard problem!**”

In Sum: Presume popular, modern view, no extreme assumption.

Premise 2: Aware AI

Aware = accounting for circumstances in a broad way, reacting to complex situations in an adequate way.

AI that makes it more obvious, also to laypeople, how complex thinking and reacting can take place in CPUs.

No genuine revolution.

In Sum: Presume the AI everyone agrees we'll develop tomorrow, not necessarily any superintelligence explosion.

Intro

Premises

- P1 Weak Computationalism
- P2 Aware AI

Hypotheses

- H1 AI ⇒ Illusionism
  - a. Reality: AI suggest Illusionism is True?
  - b. Popular: AI makes Illusionism more Popular?
- H2 Illusionism ⇒ Less Altruism?
  - a. Reality: Illusionism *Justifies* Less Altruism?
  - b. Popular: Illusionist Reduces Altruistic?

Implications

- Altruism vs. Cooperation
- Policy
- Consider Trump

Wilder Speculation

- On the Fermi Paradox
- An Altruism Paradox

1. AI ⇒ Illusionism

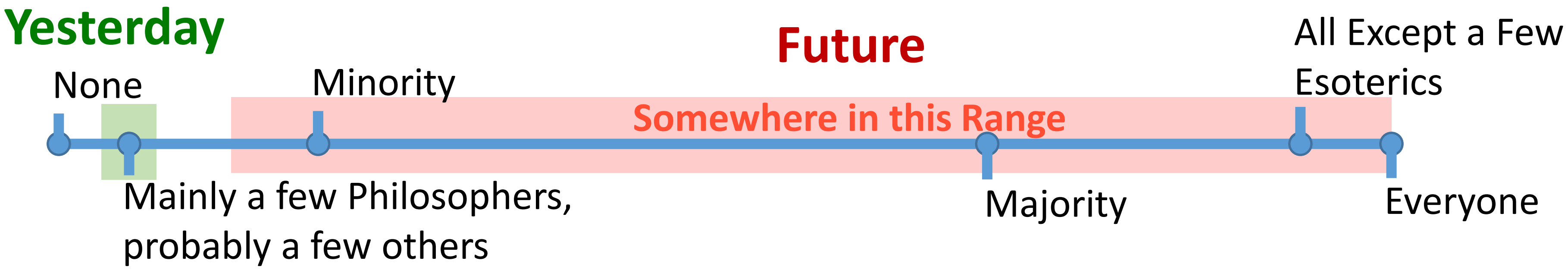
Aim is not a conclusive argument for Illusionism

Instead:

a. Why Illusionism more plausible than many think

b. The lure of Illusionism independently of whether true

Key claim: Experiencing more broadly aware AI, somewhere **between a hidden minority or nearly anyone will adopt some Illusionism**



Hypothesis 1a: Illusionism True?

Past: Two compelling reasons suggested magic rather than Computationalism

i. Only Humans have nontrivial smartness. Likely from our magical brain!

ii. We feel that we feel. This is even more trivial than Cogito Ergo Sum

Reason i: Gone! AI has removed it

Reason ii: Tougher! Next slide few reasons pro-Illusionism[1]

[1] Habermacher 2019 brought forward various of the arguments used here to argue for "Very Strong Computationalism" akin to Illusionism. Also, for example, Frankish 2016 provides a comprehensive argumentation supporting Illusionism. Similarly Chalmers 2018 explains well why Illusionism is to be taken seriously.

Hypothesis 1a: Illusionism True?

1. Hard Problem equally weird
2. God, Free Will: Many were 100%, unquestionably sure to feel X; later they or others figure, Not X. Examples: God. Free Will.
3. Weak computationalism allows to feel, but disallows to act on feelings! Hence it’d disallow you to come to this conference merely because you feel there’s a hard problem!
4. It may be easy to make a Zombie be 100% convinced he’s sentient! Maybe I can program into him really elaborate – or simple – ‘feeling convictions’.
5. Have two symmetric brains: 1 muh apart; touching; merged. When are there 2 and when 1 consciousness? Happy with any answer? (Habermacher 2019)

Hypothesis 1b: Illusionism becomes Popular?

Philosophers may provide elaborate arguments for and against the idea of a strict consciousness divide between AI and us.

No guarantee they reach laypeople, no guarantee they override beliefs.

Complex topics: Believes = based on feelings, not what is logically most salient (Haidt 2012).

Trivial topics, arguably: Believes = track the simple facts/logics.

Bots may seem very human to us → Difficult to distinguish our and their case for phenomenology

Bots being pure maths and physical electron flow: difficult to deny eventually.

➔ **Bots = computation; Humans ≈ Bots** **= View of the Illusionist!**

Update: The banal LaMDA (2022) incident is an early illustration for how quickly we some might feel HER (2013)-style connections between Humans and AI *youtu.be/sAquwhl304I*

Hypothesis 1: AI ⇒ Illusionism?

In Sum:

➔ Various elements pointing towards Illusionism becoming more popular as AI emerges

- Intro
- Premises
  - P1 Weak Computationalism
  - P2 Aware AI
- Hypotheses
  - H1 AI ⇒ Illusionism
    - a. Reality: AI suggest Illusionism is True?
    - b. Popular: AI makes Illusionism more Popular?
  - H2 Illusionism ⇒ Less Altruism?
    - a. Reality: Illusionism *Justifies* Less Altruism?
    - b. Popular: Illusionist Reduces Altruistic?
- Implications
  - Altruism vs. Cooperation
  - Policy
  - Consider Trump
  - Wilder Speculation
  - On the Fermi Paradox
  - An Altruism Paradox

Hypothesis 2a: Illusionism Justifies Less Altruism?

Utilitarianism

Probably!

Deontology / Kant

Often not

Virtue Ethics

Maybe no

But maybe yes!

Others ?

Hypothesis 2b: Illusionism Reduces Altruism?

Embryos, Animals, Computers good comparison.

In law and many people's opinion : "**Must treat nicely *iif* feel pain, phenomenologically**"

> "Animals in all likelihood seem sentient, cannot mistreat them"
>
> Abortion question often mainly about when the embryo may become sentient, conscious, pain-receptive
>
> "AI may have intrinsic value, deserve care, from the point where it becomes conscious"

Minority position: Treat animals nicely for their sake even if insentient.

Nearly unheard: "Is it evil to lose a life in Super Mario?"

Hypotheses Wrap-up

H1:

Illusionism quite some appeal

Appeal increases the more we’re exposed to Aware AI

H2:

Altruism, in practice, extremely directly depending on (believe about) others’ phenomenological experience

Absent the idea of such feeling, it may be very difficult to sustain traditional level of genuinely positive dispositions towards others.

➔ **Some or many,** in future generations, may behave **a bit or much more selfishly**. But a strong future **tendency towards egoism cannot be excluded**.

Saving Altruism?

An Illusionist need not be egoistic; she may be the kindest persons ever.

Virtue ethics and simply genuine ‘warm glow’ preferences for being kind may be enough.

Illusionist might even say ‘*I* don’t matter either, so why not simply be more altruistic instead of doing everything in my own (future) interest?!’

But the idea that others matter is a first-order reason why we care. Foolish to ignore the risk that illusionism can bear.

Saving Altruism?

Being nice to machines?

Kammerer (2019) proposes to turn implications up-side down: **elevate the moral status of (maybe even unadvanced?) machines, instead of becoming indifferent to fellow humans.**

This feels searched; more the result of the search for an answer to ‘how can we save the [pre-imposed] conclusion that we shall remain kind?’

➔ Might be a way to *rationalize ex-post* why one wants to be altruistic, rather than a logical reason *ex-ante* for why one should adopt altruism towards insentient creatures.

Kammerer rightly points out, illusionism breaks strict distinction between human and machine. **Altruism under illusionism rather concerns either your computer game too, or neither us.**

Reality: **Implausible that everyone now tries to hug their Super Mario.**

Intro

Premises

- P1 Weak Computationalism
- P2 Aware AI

Hypotheses

- H1 AI ⇒ Illusionism
    - a. Reality: AI suggest Illusionism is True?
    - b. Popular: AI makes Illusionism more Popular?
- H2 Illusionism ⇒ Less Altruism?
    - a. Reality: Illusionism *Justifies* Less Altruism?
    - b. Popular: Illusionist Reduces Altruistic?

Implications

- Altruism vs. Cooperation
- Policy
- Consider Trump

Wilder Speculation

- On the Fermi Paradox
- An Altruism Paradox

Altruism vs. Cooperation

Altruism = I care for your own sake. Or for my warm-glow from it.

Cooperation = We deal with each other for our both benefits

Some Cooperation may still be possible. But cold-hearted. Pure strategy. Or based on *conscious* warm-glow.

Policy

- Today already, modern society fraught with nearly unsolvable challenges
    - “Unfit for the Future” (Persson & Savolescu 2012, Oxford)
    - Climate change; Nuclear weapons; Tons of geopolitical challenges; Domestic inequality; Corona as example how badly we can react; Huge share of population governed by dictators without much hope for improvement. Worse, state of modern democracies may make it difficult to complain against dictators; …

- Current society cannot function without altruistic basis (Kirchgässner 2010, Sen 1977)

- Non-illusionists might have to be the ones fighting the impact
    - **Fight for policies keeping people’s egoism in check**
    - Not against illusionism, or for altruism; may be futile

Consider Trump

Many claim: Trump = Selfish, putting the US and maybe the world at stake for own gains etc.[1] *For the sake of the argument, let's assume there's something to it.*

Experiment: **What if he was Illusionist**, and consistently so?!

- **You could barely be angry at him!**
  - Yes, a zoo this world is for him, a real game.
  - Not his fault. In his mind it **genuinely matters as little as if it all was a computer game!**

- **Such people might exist**, and as put forward here, **may rapidly become common!**
  - Maybe many not very bad, but simply not so 100% serious anymore about behaving decently

➔ Harris (2020) may be righter than we think when he says "what […] has taught us beyond any possibility of doubt is that we can't rely on human decency. We need a system that can handle a psychopath in the White House."

[1] As exceptionally in'humane' as many think, or does he just have the incentive and opportunity to do it on a much more massive scale than others?

Intro

Premises

- P1 Weak Computationalism
- P2 Aware AI

Hypotheses

- H1 AI ⇒ Illusionism
  - a. Reality: AI suggest Illusionism is True?
  - b. Popular: AI makes Illusionism more Popular?
- H2 Illusionism ⇒ Less Altruism?
  - a. Reality: Illusionism *Justifies* Less Altruism?
  - b. Popular: Illusionist Reduces Altruistic?

Implications

- Altruism vs. Cooperation
- Policy
- Consider Trump

Wilder Speculation

- On the Fermi Paradox
- An Altruism Paradox

Speculation

Could it explain Fermi’s Paradox?

Fermi Paradox = Absence of evidence for extraterrestrial civilizations despite high estimates for their probability.

Speculation

Could it explain Fermi’s Paradox?

A. Retrospective version (Lucky us)

What if:

1. Smart ⇒ Lethal-at-ease (obvious)
2. Me Lethal-at-ease ⇒ You sleepless unless I’m altruistic (obvious)
3. **Illusionism**, aka my conviction that you & I both intrinsically ‘matter’, **key ingredient to reach altruism** of a kind that was evolutionarily fruitful (incl. evolutionarily stable)?
    1. Speculative
    2. Or maybe even just nearly a way of framing a tautology?!

➔ **Intelligence explosion of cooperative humans impossible without illusionism**, that weird loop in our brain claiming we really matter

➔ **Maybe that weird loop happens rarely.** After all, it is a weird backdoor for bringing in broad altruism

Speculation

Could it explain Fermi’s Paradox?

B. Forward looking version (Doom)

What if AI affecting altruism so severely, and altruism so crucial, that future society ungovernable?

Theoretically could happen each time a species is about to create superintelligence

“Artificial Superintelligence is killing cooperative society*, before it even exists!”

*and along its own emergence

Speculation

An Altruism Paradox

- Evolution as individuals within competing groups has made us altruists to a degree, say x% on scale from 0 to 100%
- Evolved as species/society collaborating and thriving more or less well
- Rules, control mechanisms, technology, norms evolved, stabilizing society against egoistic traits
  - Thieves, speedsters on the road, rapists, …, tend to be reasonably contained
- Imagine x was bit higher. Society would have functioned better until now.
- If Illusionism about to emerge, the most altruistic society may be the most vulnerable:

As a species, best chance to survive illusionism shift if was most egoistic before!

Cannot be excluded that worst offender may have been the best thing to happen to society!

Conclusions

1. AI may make Illusionism more prominent, for good or for bad
2. Illusionism puts Altruism at risk
    1. Few slightly more egoistic people may have a severe impact on society
    2. Many much more egoistic people are even possible
3. Any single person truly not caring about society because of her illusionism may be a serious danger
4. Policies needed, to better contain egoistic behavior, save public goods

Speculation

5. *Fermi Paradox 1: Maybe illusionism was required for us to develop?*
6. *Fermi Paradox 2: Maybe overcoming illusionism will be our downfall?*
7. *Altruism Paradox: Might the most egoistic society be most robust to suddenly emerging Illusionism?*

Thank you

Florian Habermacher

florian.habermacher@gmail.com

Views are Author’s own, unrelated to affiliations.

References

- Bostrom 2006 Quantity of experience: Brain-duplication and degrees of consciousness, Mind and Machines
- Café Phi 2019 François Kammerer - Sommes-nous tous des zombies ?
- Chalmers 1995 Moving Forward on the Problem of Consciousness
- Chalmers 2018 The Meta-Problem of Consciousness
- Frankish Keith 2019 Illusionism as a Theory of Consciousness
- Habermacher 2019 The Case for Very Strong Computationalism - Is Phenomenological Consciousness, the Hard problem merely Logical State + Own Conviction to Matter + Semantic confusion? TSC Conference 2019
- Haidt 2012 The Righteous Mind
- S. Harris 2020 Making Sense: Republic of Lies (episode 225) (transcript: https://www.happyscribe.com/public/making-sense-with-sam-harris/225-republic-of-lies)
- HER Movie 2013. Clip
- Kirchgässner 2010 On Minimal Morals, Europ Journ of Polit Econ
- LaMDA 2022, Washington Post: The Google engineer who thinks the company’s AI has come to life
- Persson & Savolescu 2012 Unfit for the Future: The Need for Moral Enhancement, Oxford Uni Press
- Sen 1977 Rational Fools: a critique of the behavioral foundations of economic theory. Philosophy and Public Affairs