Artificial Moral Advisor and Moral Enhancement

Yuxin Liu, Adam Moore, & Matti Wilks

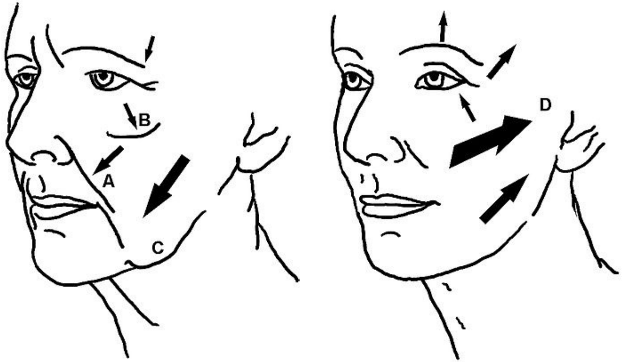
AI Ethics and Human-Computer Interaction Conference 2024

March 07, 2024
Graz, Austria

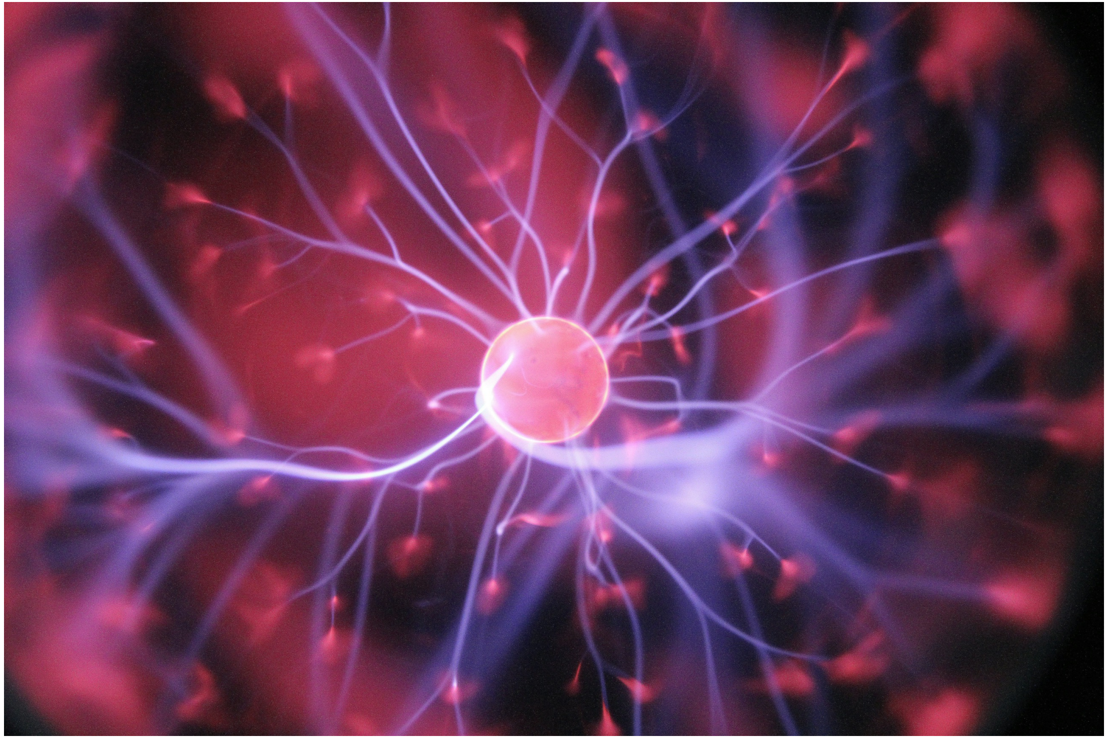
Centre for **Technomoral Futures**

THE UNIVERSITY *of* EDINBURGH
School of Philosophy, Psychology
and Language Sciences

Human Enhancement

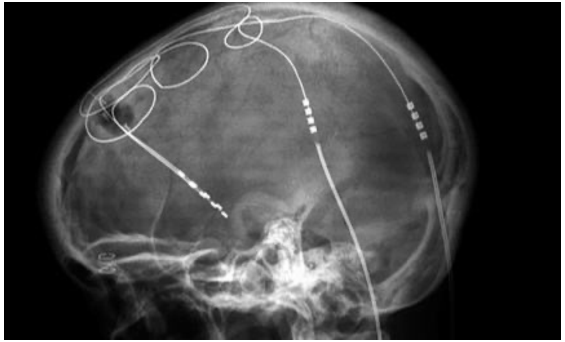


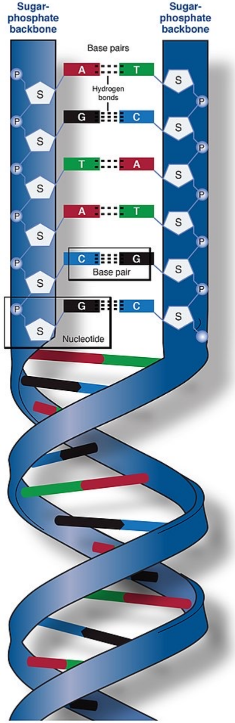
Biocognitive Enhancement

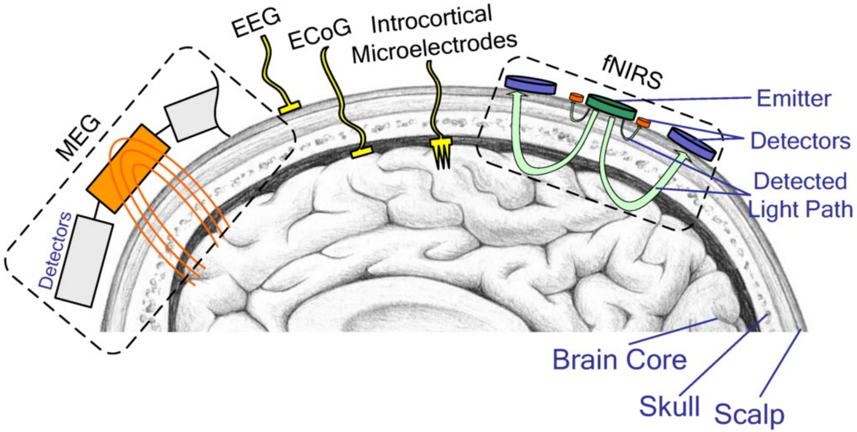


Moral Enhancement

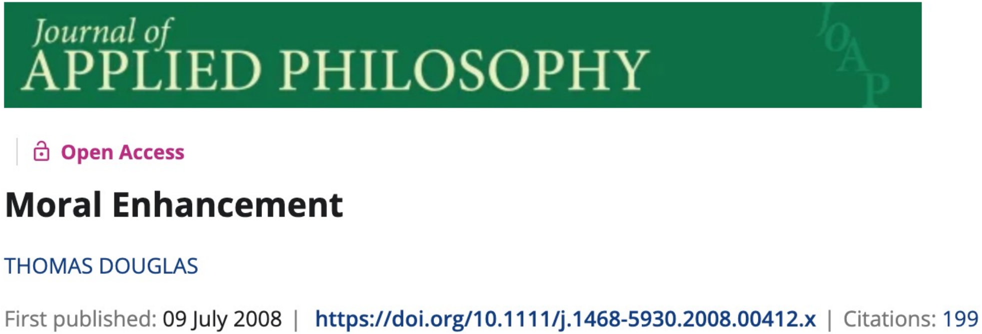
Proposed Methods







Moral (Bio)enhancement



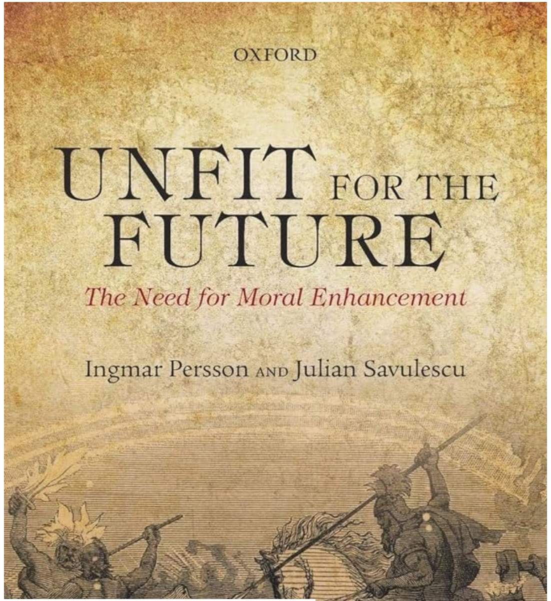
Journal of APPLIED PHILOSOPHY

Open Access

Moral Enhancement

THOMAS DOUGLAS

First published: 09 July 2008 | https://doi.org/10.1111/j.1468-5930.2008.00412.x | Citations: 199

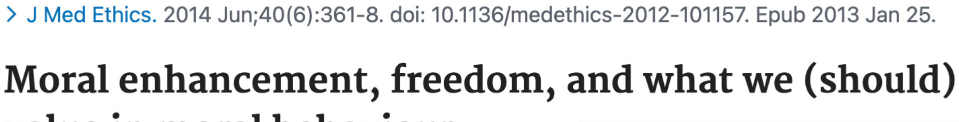


OXFORD

UNFIT FOR THE FUTURE

The Need for Moral Enhancement

Ingmar Persson AND Julian Savulescu

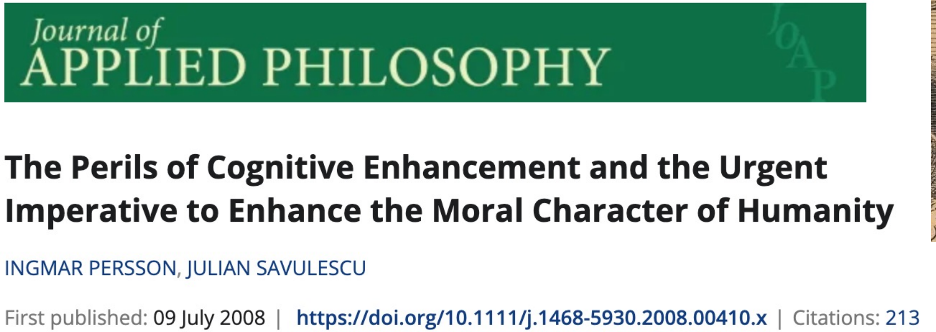


J Med Ethics. 2014 Jun;40(6):361-8. doi: 10.1136/medethics-2012-101157. Epub 2013 Jan 25.

Moral enhancement, freedom, and what we (should) value in moral behaviour

David DeGrazia

PMID: 23355049 DOI: 10.1136/medethics-2012-101157



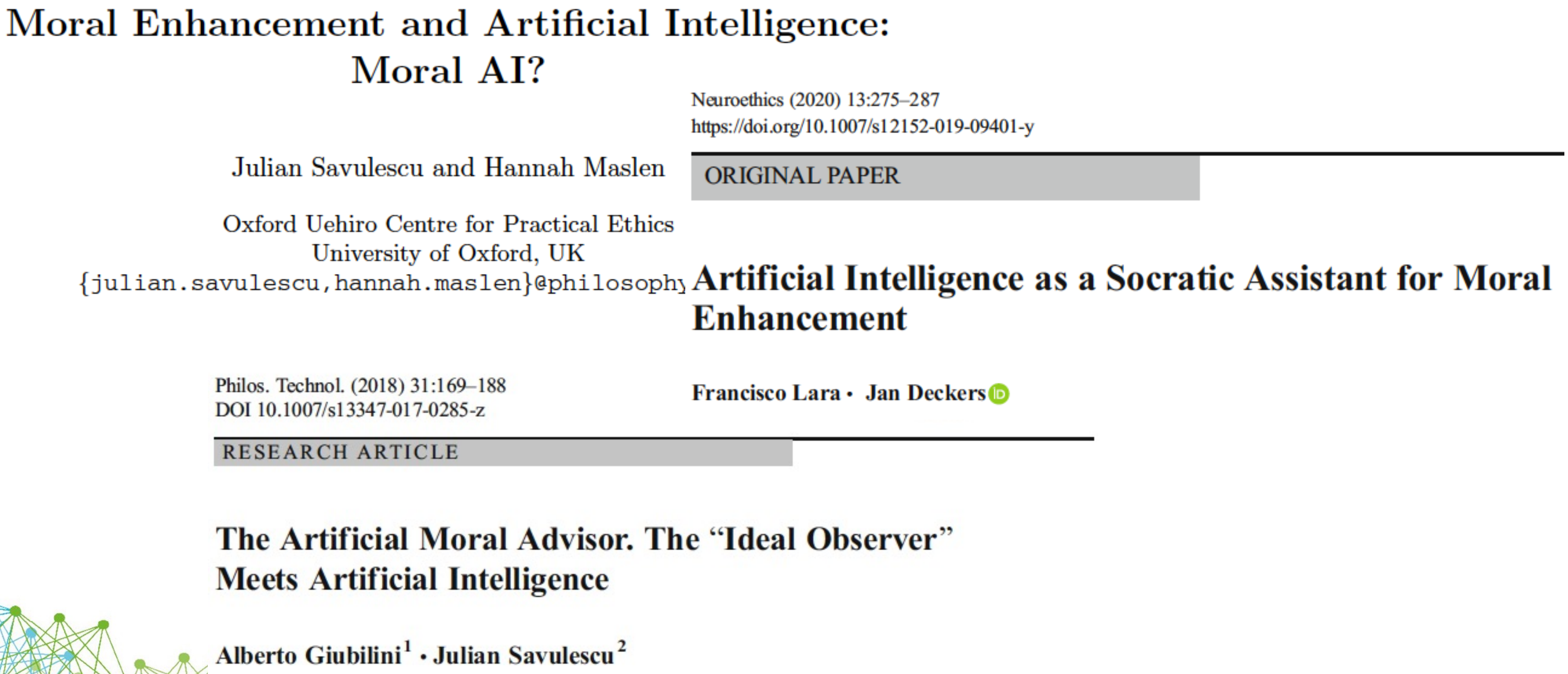
Journal of APPLIED PHILOSOPHY

The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity

INGMAR PERSSON, JULIAN SAVULESCU

First published: 09 July 2008 | https://doi.org/10.1111/j.1468-5930.2008.00410.x | Citations: 213

Moral AI Enhancement



Moral Enhancement and Artificial Intelligence: Moral AI?

Julian Savulescu and Hannah Maslen

Oxford Uehiro Centre for Practical Ethics
University of Oxford, UK
{julian.savulescu,hannah.maslen}@philosophy

Neuroethics (2020) 13:275–287
https://doi.org/10.1007/s12152-019-09401-y

ORIGINAL PAPER

Artificial Intelligence as a Socratic Assistant for Moral Enhancement

Francisco Lara • Jan Deckers

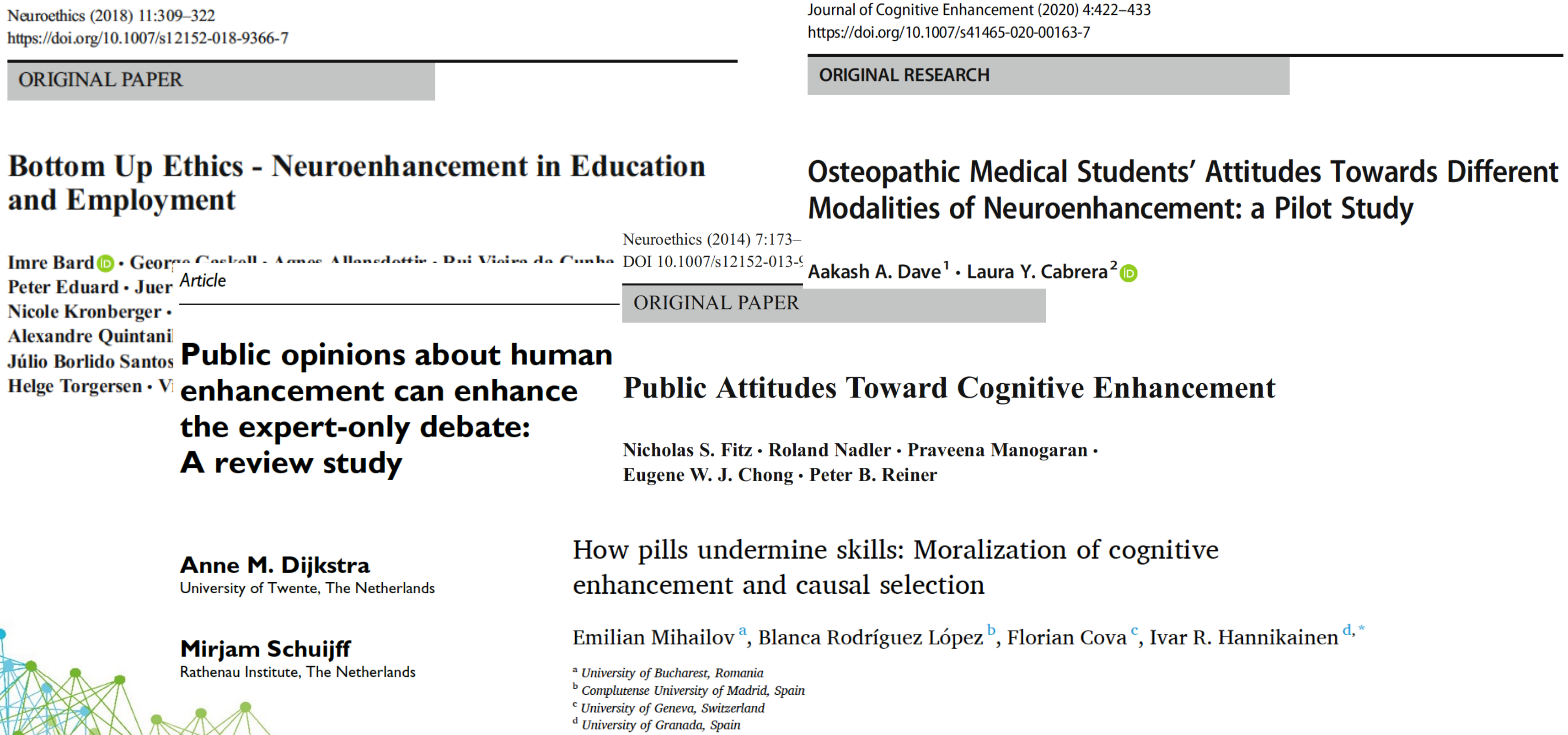
Philos. Technol. (2018) 31:169–188
DOI 10.1007/s13347-017-0285-z

RESEARCH ARTICLE

The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence

Alberto Giubilini[1] • Julian Savulescu[2]

Public Perceptions of Enhancement



Neuroethics (2018) 11:309–322
https://doi.org/10.1007/s12152-018-9366-7

ORIGINAL PAPER

Bottom Up Ethics - Neuroenhancement in Education and Employment

Imre Bard · George Gaskell · Agnes Allansdottir · Rui Vieira da Cunha
Peter Eduard · Juer
Nicole Kronberger ·
Alexandre Quintanil
Júlio Borlido Santos
Helge Torgersen · Vi

Journal of Cognitive Enhancement (2020) 4:422–433
https://doi.org/10.1007/s41465-020-00163-7

ORIGINAL RESEARCH

Osteopathic Medical Students' Attitudes Towards Different Modalities of Neuroenhancement: a Pilot Study

Aakash A. Dave[1] · Laura Y. Cabrera[2]

Article

Public opinions about human enhancement can enhance the expert-only debate: A review study

Anne M. Dijkstra
University of Twente, The Netherlands

Mirjam Schuijff
Rathenau Institute, The Netherlands

Neuroethics (2014) 7:173–
DOI 10.1007/s12152-013-

ORIGINAL PAPER

Public Attitudes Toward Cognitive Enhancement

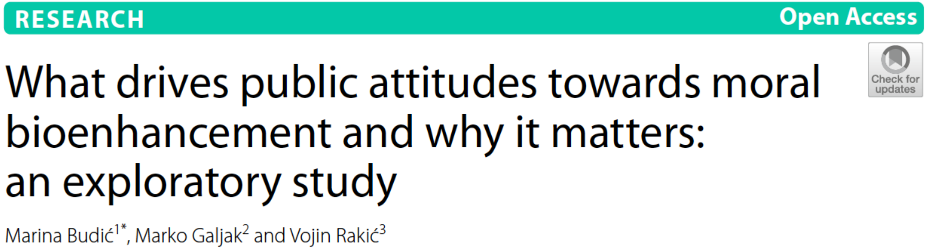
Nicholas S. Fitz · Roland Nadler · Praveena Manogaran ·
Eugene W. J. Chong · Peter B. Reiner

How pills undermine skills: Moralization of cognitive enhancement and causal selection

Emilian Mihailov[a], Blanca Rodríguez López[b], Florian Cova[c], Ivar R. Hannikainen[d,*]

[a] University of Bucharest, Romania
[b] Complutense University of Madrid, Spain
[c] University of Geneva, Switzerland
[d] University of Granada, Spain

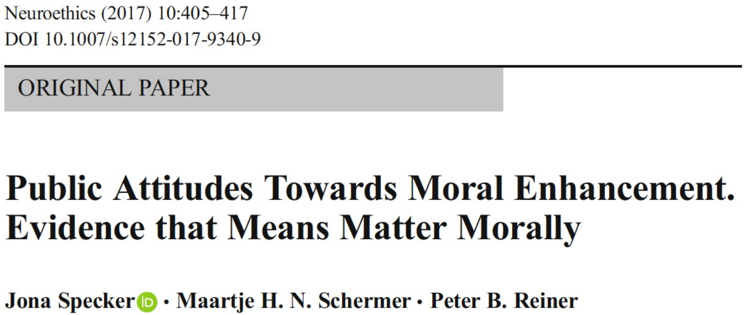
Public Perceptions of Moral Enhancement



RESEARCH Open Access

What drives public attitudes towards moral bioenhancement and why it matters: an exploratory study

Marina Budić[1*], Marko Galjak[2] and Vojin Rakić[3]



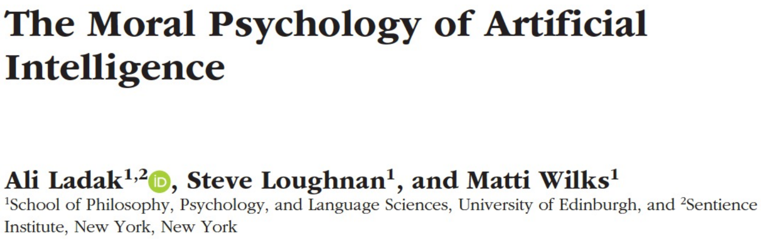
Neuroethics (2017) 10:405–417
DOI 10.1007/s12152-017-9340-9

ORIGINAL PAPER

Public Attitudes Towards Moral Enhancement. Evidence that Means Matter Morally

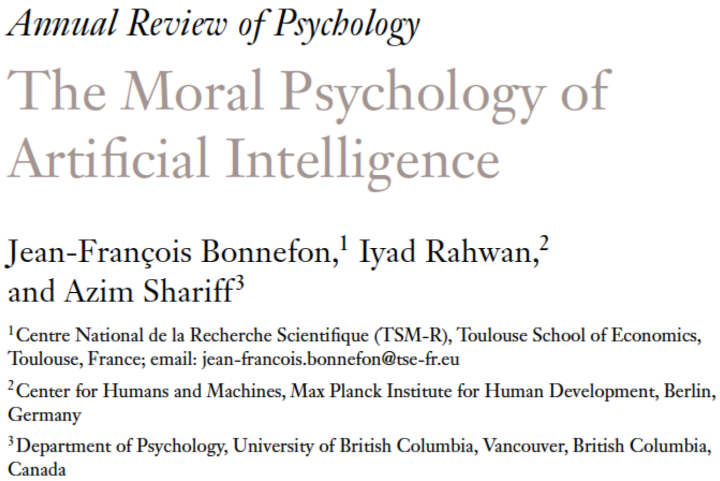
Jona Specker · Maartje H. N. Schermer · Peter B. Reiner

Moral Psychology of AI



The Moral Psychology of Artificial Intelligence

Ali Ladak[1,2], Steve Loughnan[1], and Matti Wilks[1]
[1]School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, and [2]Sentience Institute, New York, New York



Annual Review of Psychology

The Moral Psychology of Artificial Intelligence

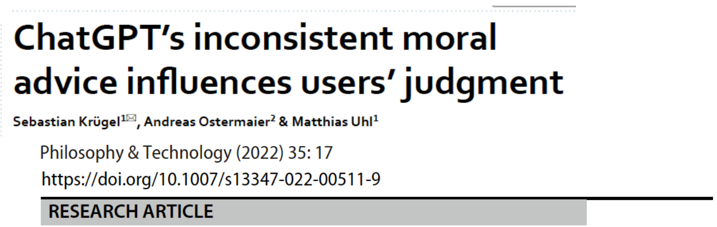
Jean-François Bonnefon,[1] Iyad Rahwan,[2] and Azim Shariff[3]

[1]Centre National de la Recherche Scientifique (TSM-R), Toulouse School of Economics, Toulouse, France; email: jean-francois.bonnefon@tse-fr.eu

[2]Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany

[3]Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada

Moral Psychology of AI

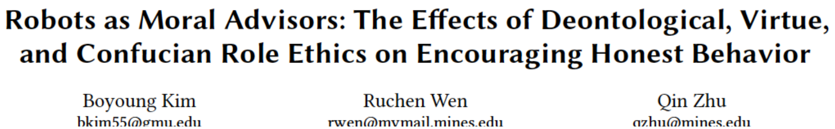


ChatGPT's inconsistent moral advice influences users' judgment

Sebastian Krügel[1]✉, Andreas Ostermaier[2] & Matthias Uhl[1]

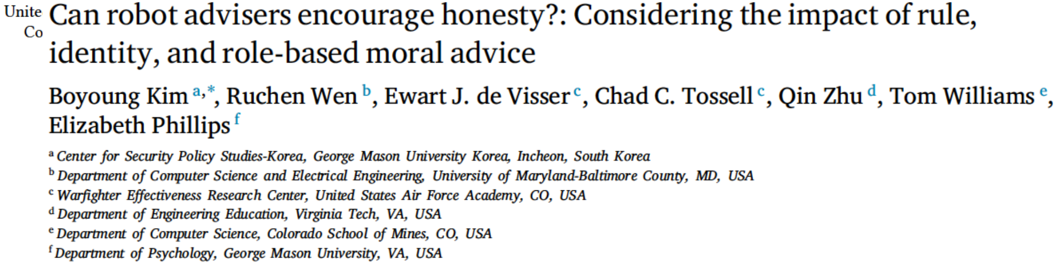
Philosophy & Technology (2022) 35: 17
https://doi.org/10.1007/s13347-022-00511-9

RESEARCH ARTICLE



Robots as Moral Advisors: The Effects of Deontological, Virtue, and Confucian Role Ethics on Encouraging Honest Behavior

Boyoung Kim bkim55@gmu.edu
Ruchen Wen rwen@mymail.mines.edu
Qin Zhu qzhu@mines.edu



Unite
Co

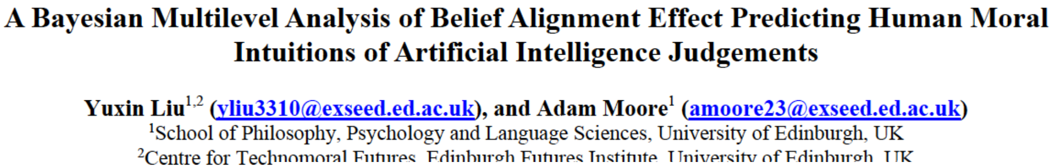
Can robot advisers encourage honesty?: Considering the impact of rule, identity, and role-based moral advice

Boyoung Kim[a,*], Ruchen Wen[b], Ewart J. de Visser[c], Chad C. Tossell[c], Qin Zhu[d], Tom Williams[e], Elizabeth Phillips[f]

[a] *Center for Security Policy Studies-Korea, George Mason University Korea, Incheon, South Korea*
[b] *Department of Computer Science and Electrical Engineering, University of Maryland-Baltimore County, MD, USA*
[c] *Warfighter Effectiveness Research Center, United States Air Force Academy, CO, USA*
[d] *Department of Engineering Education, Virginia Tech, VA, USA*
[e] *Department of Computer Science, Colorado School of Mines, CO, USA*
[f] *Department of Psychology, George Mason University, VA, USA*

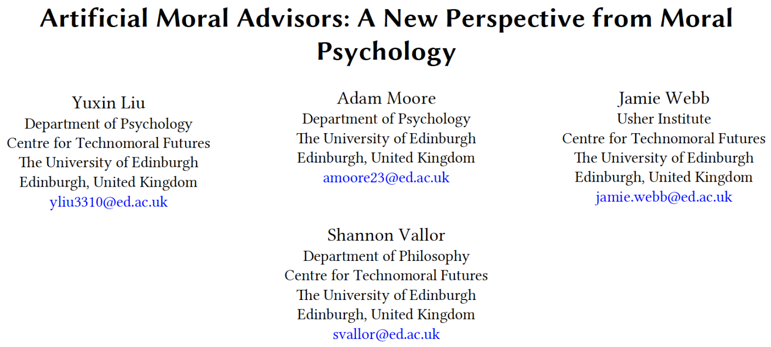
Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions

Sebastian Krügel[1,2] · Andreas Ostermaier[3] · Matthias Uhl[1]



A Bayesian Multilevel Analysis of Belief Alignment Effect Predicting Human Moral Intuitions of Artificial Intelligence Judgements

Yuxin Liu[1,2] (yliu3310@exseed.ed.ac.uk), and Adam Moore[1] (amoore23@exseed.ed.ac.uk)
[1]School of Philosophy, Psychology and Language Sciences, University of Edinburgh, UK
[2]Centre for Technomoral Futures, Edinburgh Futures Institute, University of Edinburgh, UK



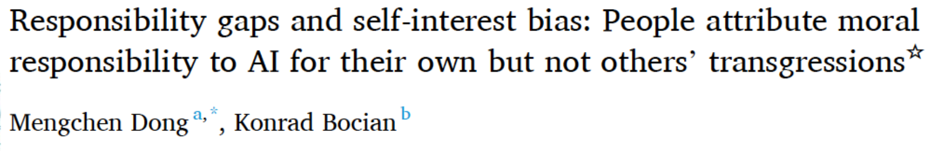
Artificial Moral Advisors: A New Perspective from Moral Psychology

Yuxin Liu
Department of Psychology
Centre for Technomoral Futures
The University of Edinburgh
Edinburgh, United Kingdom
yliu3310@ed.ac.uk

Adam Moore
Department of Psychology
The University of Edinburgh
Edinburgh, United Kingdom
amoore23@ed.ac.uk

Jamie Webb
Usher Institute
Centre for Technomoral Futures
The University of Edinburgh
Edinburgh, United Kingdom
jamie.webb@ed.ac.uk

Shannon Vallor
Department of Philosophy
Centre for Technomoral Futures
The University of Edinburgh
Edinburgh, United Kingdom
svallor@ed.ac.uk



Responsibility gaps and self-interest bias: People attribute moral responsibility to AI for their own but not others' transgressions☆

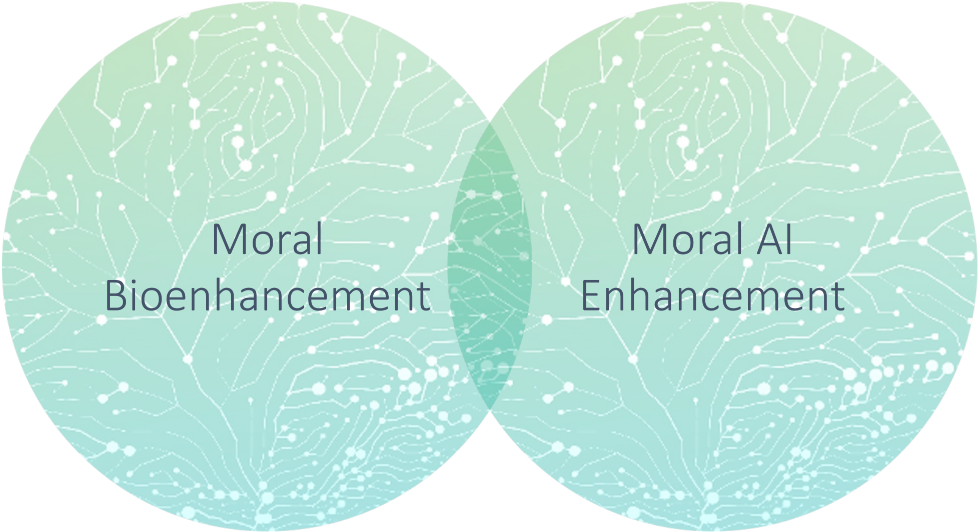
Mengchen Dong[a,*], Konrad Bocian[b]

[a] *Center for Humans and Machi*
[b] *Department of Psychology in S*

When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions

Daniel B. Shank, Alyssa DeSanti and Timothy Maninger

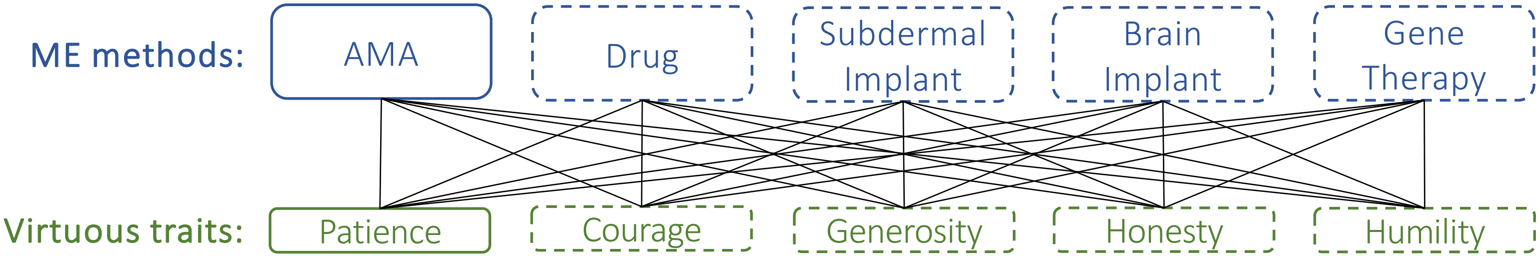
The Current Study



1. Biological Alteration
2. Violation of Essentialism
3. Effort-Saving
4. Efficacy
5. Familiarity

PREREGISTERED

Experimental Design



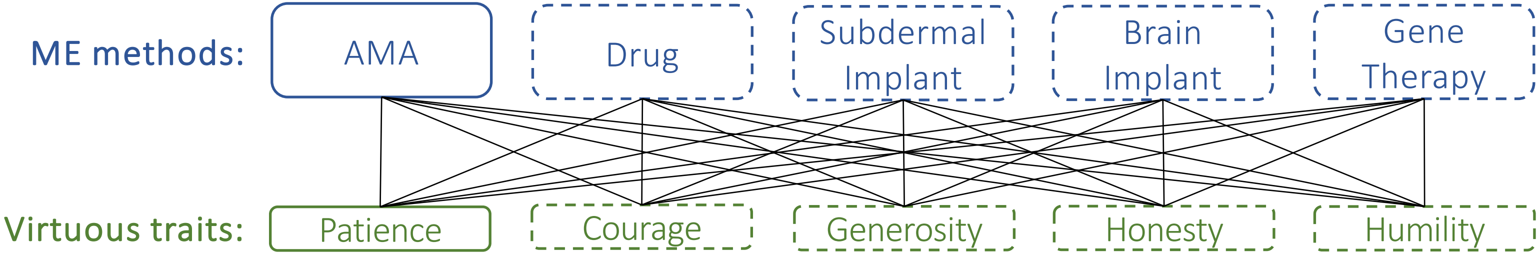
Experimental Design

PREREGISTERED

In the near future, significant achievements in science have opened up the possibility of moral enhancement — the ability to improve people's moral characteristics through a range of medical or non-medical technologies. Evidence has shown that these technologies are safe for humans. People who want to morally improve themselves can voluntarily choose to make use of these moral enhancement technologies. One of these technologies is an ***artificial moral advisor (AMA)****, which can effectively enhance one's moral capacity through an external AI device that produces moral advice based on signals from information in one's surrounding physical environment. For example, it can help a person focus on their own emotions in the moment and their underlying causes, so that they are more likely to be* ***patient****. Imagine a scenario where Sam has been working on an important report with a co-worker who fails to deliver their part after a whole week. Before enhancement, Sam would have lashed out. After using this technology, Sam is easily able to consider various possible external reasons for the co-worker's delay.*

Experimental Design

PREREGISTERED



ME methods: AMA, Drug, Subdermal Implant, Brain Implant, Gene Therapy

Virtuous traits: Patience, Courage, Generosity, Honesty, Humility

IVs

1. Biological Alteration
2. Violation of Essentialism
3. Effort-Saving
4. Efficacy
5. Familiarity

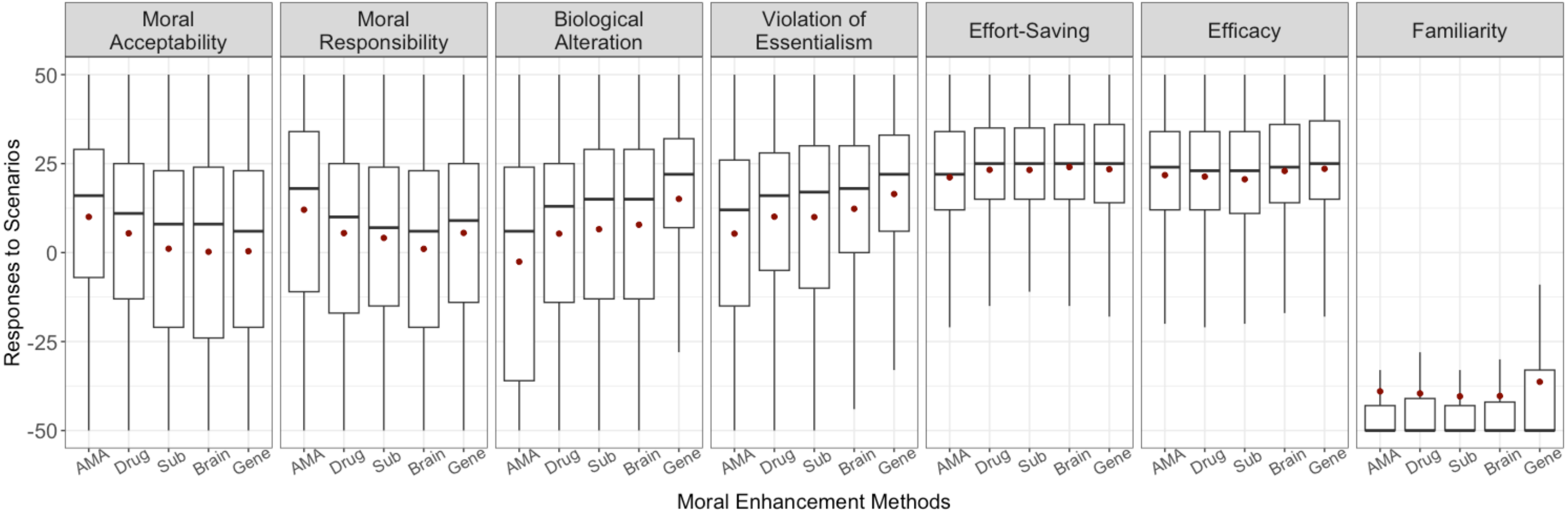
(all interacting with method)

DVs

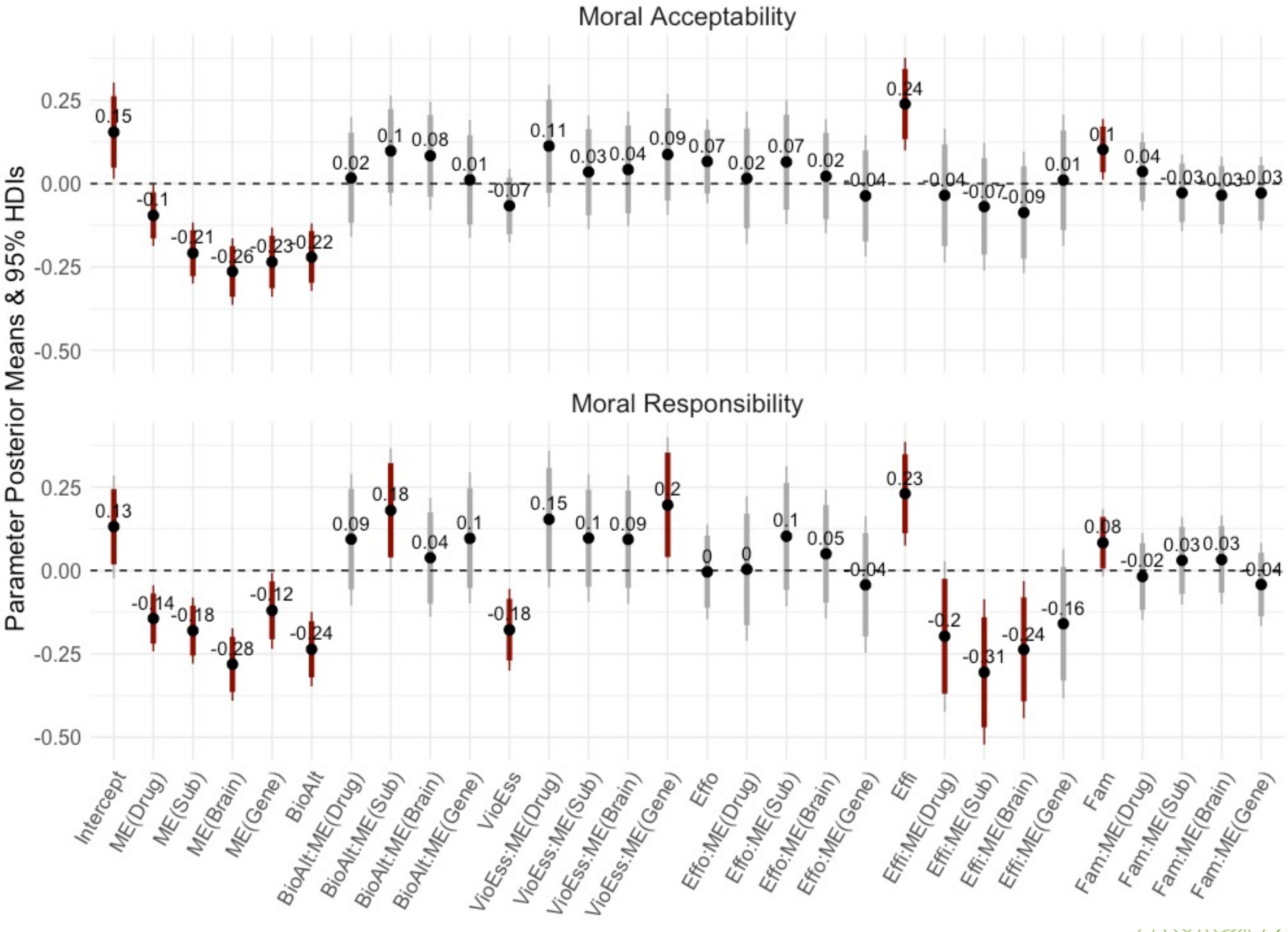
1. Moral acceptability
2. Moral responsibility

Results

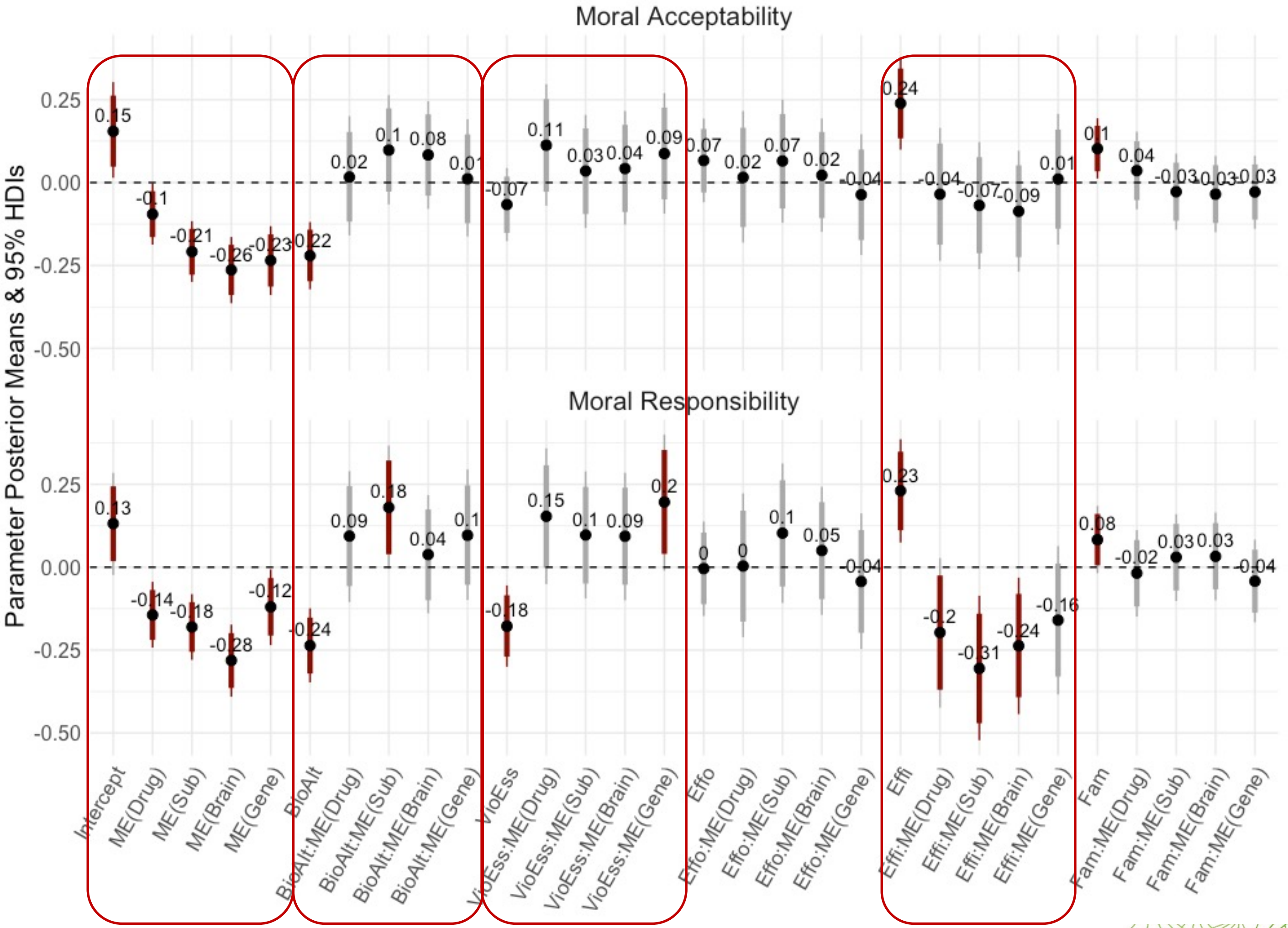
301 UK subjects via Prolific (192 females & 108 males; *M* = 39.09 yrs, *SD* = 10.85 yrs)



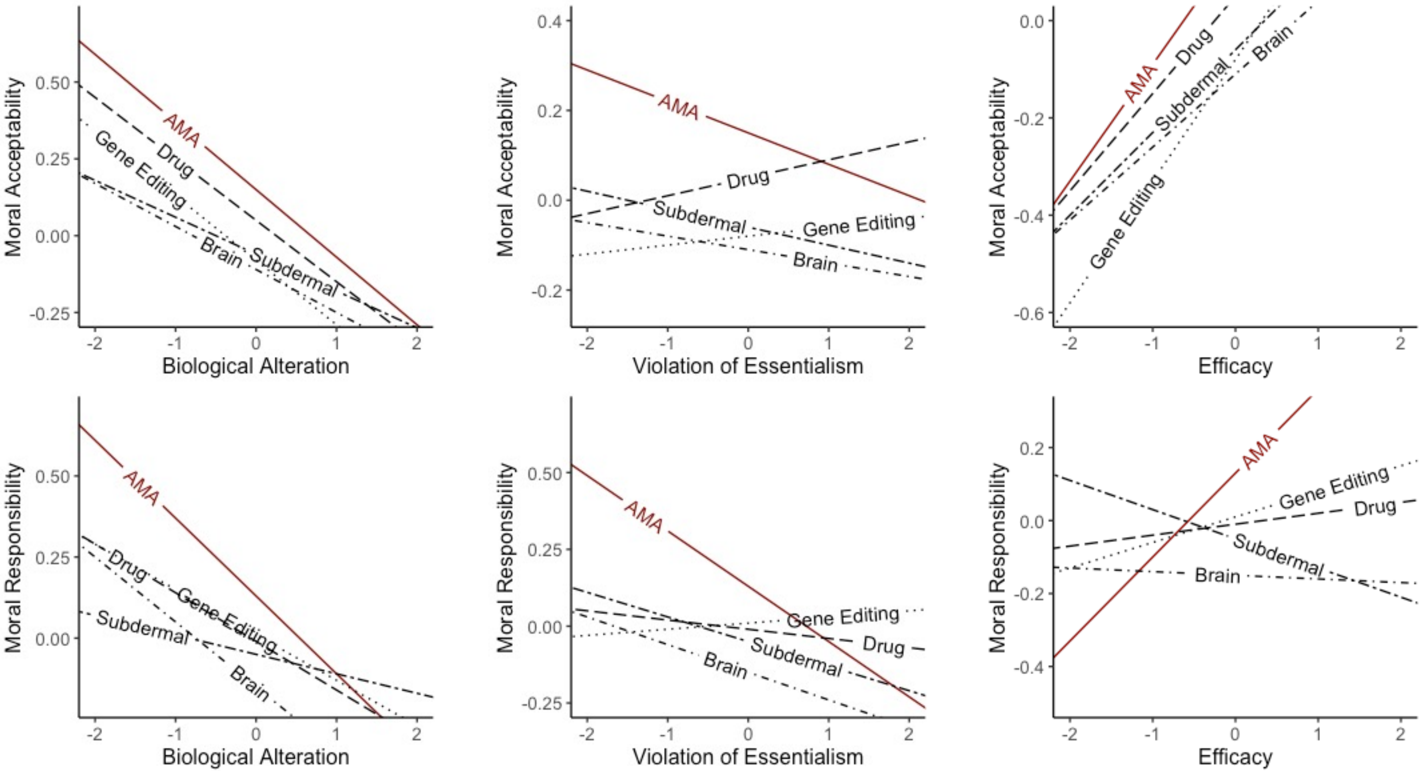
Descriptive summary of main variables grouped by moral enhancement method



Bayesian multilevel multivariate multiple regression



Bayesian multilevel multivariate multiple regression



Regression model interaction effects

Summary

AMA results in higher moral acceptability and greater attribution of moral responsibility than moral bioenhancements

Moral acceptability declines with greater biological alteration, but can be improved with greater efficacy

- Practical implication for adoption of moral enhancement

Moral responsibility for the AMA-enhanced is reduced with more changes to biology/human nature, and increases with greater perceived efficacy; but this shift is diminished for the biomedically-enhanced, such that they are still almost entirely self-responsible

- Risks of AI-scapegoating?

Centre for Technomoral Futures

technomoralfutures.uk/phd-fellows/yuxin-liu

yliu3310@ed.ac.uk

yliu-psych.github.io

@_yuxinl_

Thank you!