

# ETHICS of AI/LLMs inc ChatGPT

**Professor Julian Savulescu**

Chen Su Lan Centennial Professor

Director Centre for Biomedical Ethics

Yong Loo Lin School of Medicine

National University of Singapore

FOR SUBSCRIBERS

## *The Petrov dilemma: Moral responsibility in the age of ChatGPT*

The way a Soviet officer dealt with potential nuclear armageddon holds an important lesson in how we deal with the myriad challenges thrown up by ChatGPT and others of its kind

Julian Savulescu



ChatGPT

# Messages

- Think First
- Take Responsibility
- Act Ethically

# Ethics of AI

- Ethical Justification, **NOT** Explanation
  - We don't care *why* Petrov acted (fear, diffidence, doubt, etc)
  - We care how he acted – what the result was

# What Matters in Medicine

- Justice – fair distribution of benefits and burdens
  - Beneficence – promoting interests or well-being
  - Autonomy
- 
- What matters is whether the outcome can be justified in these terms, not how or why the outcome arose
  - Black box or interpretable? – only matters in relation to present and future outcomes

# Basic Message: Ethics of AI

- Science: what humans + AI (“network” – Fei Song) are doing, what are the effects of their hybrid action
  - Randomized controlled trials
  - Ongoing ecological surveillance, auditing
- Ethics: to evaluate the outcome
  - Justification, Not Explanation
  - Trust is built on reliability and justifiability
- Science + Ethics

# Ethics of LLM inc ChatGPT

– **Credit: Sebastian Porsdam Mann, Brian Earp**

# Current Controversies in LLMs/ChatGPT in Medicine

1. Personalized LLM
2. Cocreation
3. Praise/Blame asymmetry
4. Consent
5. Ethical Avatar/P4
6. Moral Guru
7. Models of Doctor-Patient Relationship



# Current Controversies in LLMs/ChatGPT

## **1. *Personalized LLM***

2. Cocreation

3. Praise/Blame asymmetry

4. Consent

5. Ethical Avatar/P4

6. Moral Guru

7. Models of Doctor-Patient Relationship


*American Journal of Bioethics* (in press)

Article

Full-text available

**AUTOGEN: A Personalized Large Language Model for Academic Enhancement – Ethics and Proof of Principle**

December 2023 · [The American Journal of Bioethics](#)

Sebastian Porsdam Mann ·  Brian D. Earp · Nikolaj Møller · [Show all 5 authors](#) · Julian Savulescu

# Fine-tuning

- Prompt template:

Imagine that you are an academic writing a research paper.

The paper should be as interesting, comprehensive, clear, and concise as possible.

Based on the below title and abstract, write the section on “[section X]”.

Title: [Title].

Abstract: [Abstract].

Section:

# The models

Model	Nr. of publications	Nr. of prompt-completion pairs	Estimated nr. of tokens	Nr. of training epochs	Fine-tuning costs
SPM1	8	101	101k	4	\$12.24
SPM2	8	101	68k	4	\$8.17
SPM3	8	101	68k	12	\$24.52
BDE1	65	569	679k	4	\$81.56
BDE2	65	569	529k	4	\$63.55
JS	73	656	647k	4	\$77.73
CO	146	1326	1452k	4	\$174.53



Home Insert Draw Page Layout Formulas Data Review View Automate Tell me
Comments Share
Paste Arial (Body) 10 Bold Italic Underline Merge & Centre General Conditional Formatting Format as Table Cell Styles Insert Delete Format Sort & Filter Find & Select Analyse Data Enquiries

H1336 fx

Table with columns A through S and rows 1290 through 1328. The text in the cells is a continuous paragraph discussing genetic enhancement, ethical considerations, and the role of medicine.

# Current Controversies in LLMs/ChatGPT

1. Personalized LLM
- 2. *Cocreation***
3. Praise/Blame asymmetry
4. Consent
5. Ethical Avatar/P4
6. Moral Guru
7. Models of Doctor-Patient Relationship



# Example: Co-Creation

Base Model	AUTOGEN-JS
<p>As technology advances, artificial intelligence provides humanity with the opportunity to enhance their capacities and broaden their worldviews.</p>	<p>The advent of artificial intelligence, particularly large-scale language models (LLMs) such as ChatGPT, heralds a potentially transformative era for human society. However, the path to harnessing their vast potential is fraught with significant ethical and existential debates. The primary concern centres on how these advanced tools can add value to human life without compromising human dignity and authenticity. It is within this frame of discourse that this paper explores the ethical dimensions of LLM utilization, drawing parallels with the ongoing discussions on human enhancement.</p>
<p>In this paper, I will explore the value of large language models (LLMs) as a form of human enhancement, looking particularly at the two models of usage: the Slave Model and the Cocreation Model.</p>	<p>We posit that human life, in its essence, strives to generate value both for the individual and the wider society. In this context, human enhancement—through any available means, including AI—provides a unique opportunity to augment our capacity for value creation. LLMs, by virtue of their capability to synthesize vast amounts of human thought and achievement and potentially generate novel ideas, present themselves as potent tools in this endeavour.</p>
<p>By discussing the potential implications of these two models, the paper will investigate how the use of LLMs can add value in terms of pre-existing knowledge and generate new ideas, and thus provide an opportunity to authentically progress human achievement.</p>	<p>Yet, the real potential of these tools lies not just in their intrinsic capabilities, but in how we, as humans, choose to interact and engage with them. In this vein, we identify two primary modes of engagement with LLMs: the 'Slave Model' and the 'Cocreation Model'. In the former, the human is reduced to an instrument, merely transmitting or plagiarising the outputs of LLMs, thereby alienating the individual from the creative process and reducing their dignity. The Cocreation Model, conversely, calls for an active human engagement with the LLM outputs, involving their critical evaluation and elaboration, transforming the user from a passive consumer into an active participant.</p>
<p>It will ultimately consider how the use of LLMs either expand possibilities for human enhancement or risk alienating and undermining human dignity.</p>	<p>This paper argues that when used in the Cocreation Model, LLMs hold the promise of enabling authentic human progress and value creation, akin to the transformative impact of computers and the</p>

# Cocreation and Rapid proto-typing of ideas

- Can this be used to turn an idea into a workable, rough draft?
- Mandatory moral bioenhancement paper: Vaccination Analogy
  - Estimated time to write title, abstract, section headings: 15 minutes
  - Estimated time to complete material generation: 2 hours (AUTOGEN)
  - Estimated time to complete editing, structuring, formatting: 2 hours (GPT-4)
  - Total time to submission: 4 hours, 15 minutes.
  - BUT Requires idea, and this was particularly straightforward idea.
  - Also, point would be to use this material as base, not to submit.



# Rapid proto-typing of ideas: I

Title: When Should Moral Bioenhancement Be Mandatory?

Abstract:

Moral bioenhancement is the use of biomedical interventions to increase the probability of moral behaviour. This has been proposed to address existential problems such as misuse of powerful technology, such as bioterrorism, or global collective action problems such as combatting climate change. One objection is that such enhancement would only be effective if made mandatory. Yet coercion undermines freedom and autonomy, and some writers argue that moral enhancement can only truly occur if there is “freedom to fall.” We argue that coercion can be legitimately employed to promote public goods and use the parallel field of public health ethics to generate principles of coercion in the employment of moral bioenhancement. Specifically, we use an algorithm for mandatory vaccination as a basis for mandating moral bioenhancement. Mandatory moral bioenhancement can be employed when: (1) there is a grave social problem; (2) it is safe and effective; (3) its comparative expected utility is greater than less coercive options, including incentives; (4) the costs of the coercion are proportionate to the benefits (and it consists an “easy rescue” for the coerced individual).

Section headings:

Section 1. The Problem of Moral Bioenhancement

In this section we outline the concept of moral bioenhancement and major objections to it, particularly that it requires coercion if it is to be effective and this renders it no longer moral enhancement

Section 2. Public Health Ethics and Coercion

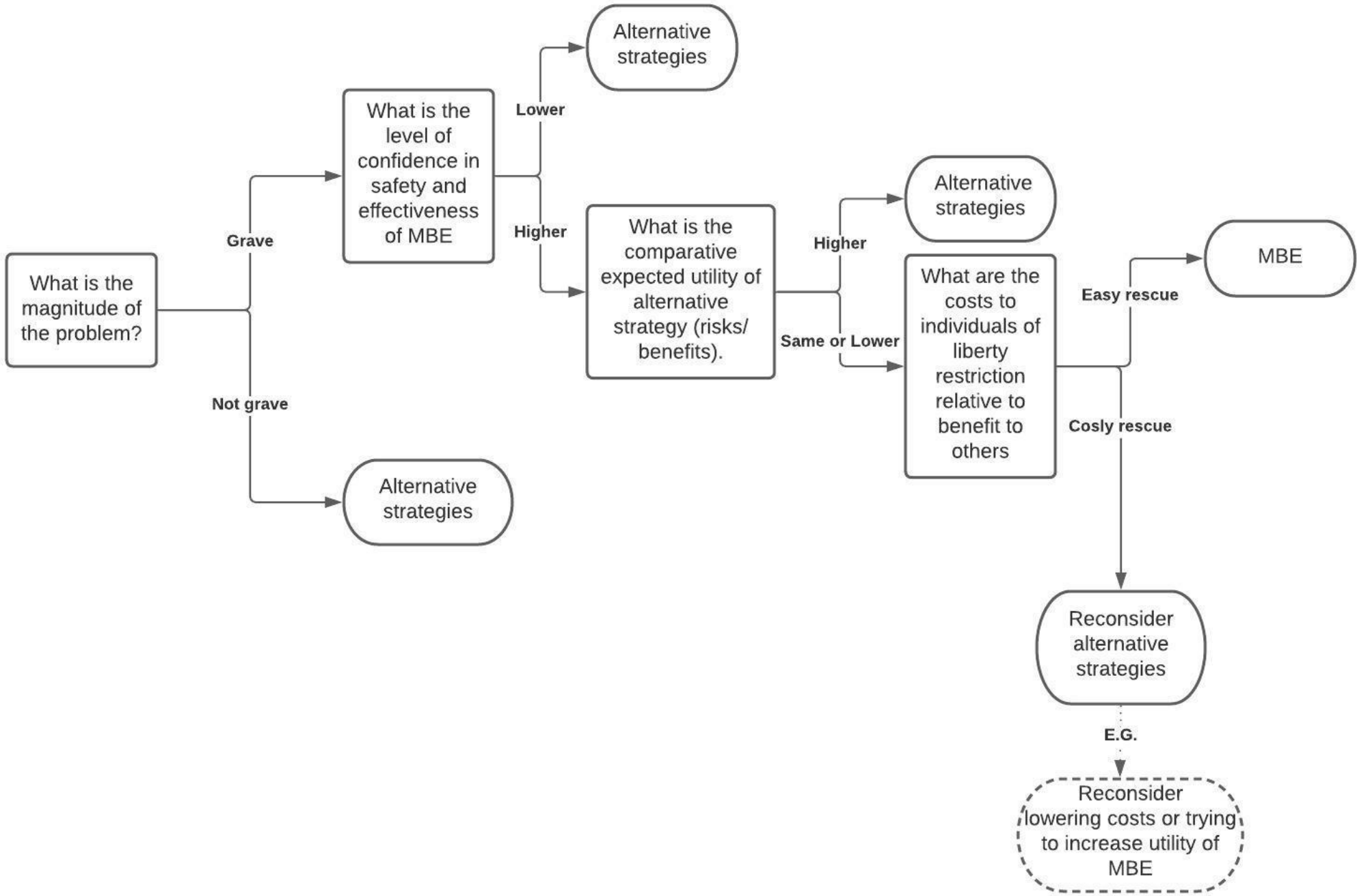
We outline the principles of when coercion can be employed to promote public health. In particular, we examine when mandatory vaccination is justified.

Section 3. Applying Public Health Ethics to Moral Bioenhancement

In this section we apply an algorithm for mandatory vaccination to mandatory bioenhancement. We outline four conditions which must be met for mandatory moral bioenhancement to be justified

Section 4. Objections

In this section, we consider objections to whether mandatory moral bioenhancement is really a moral enhancement. We argue that even if coercion exists, moral bioenhancement will remain genuinely moral if a person wishes to morally enhance themselves for the right reasons. And even in cases where a person obeys the mandate out of self-interest, we argue that costs of such an amoral enhancement may be outweighed by the benefits




# Rapid proto-typing of ideas: II

[Overview](#)[Documentation](#)[API reference](#)[Examples](#)[Playground](#)[Help](#)[Personal](#)

## Playground

[Save](#)[View code](#)[Share](#)

Imagine that you are an academic writing a paper. The paper should be as interesting, comprehensive, clear, and concise as possible. Based on the below title and abstract, write the introduction. 

Title: When Should Moral Bioenhancement Be Mandatory?

Abstract:

Moral bioenhancement is the use of biomedical interventions to increase the probability of moral behaviour. This has been proposed to address existential problems such as misuse of powerful technology, such as bioterrorism, or global collective action problems such as combatting climate change. One objection is that such enhancement would only be effective if made mandatory. Yet coercion undermines freedom and autonomy, and some writers argue that moral enhancement can only truly occur if there is "freedom to fall." We argue that coercion can be legitimately employed to promote public goods and use the parallel field of public ethics ethics to generate principles of coercion in the employment of moral bioenhancement. Specifically, we use an algorithm for mandatory vaccination as a basis for mandating moral bioenhancement. Mandatory moral bioenhancement can be employed when: (1) there is a grave social problem; (2) it is safe and effective; (3) its comparative expected utility is greater than less coercive options, including incentives; (4) the costs of the coercion are proportionate to the benefits (and it consists an "easy rescue" for the coerced individual).

Introduction: **When Should Moral Bioenhancement Be Mandatory?**

**Many thinkers on moral bioenhancement (henceforth MBI) view the safe use of such technology as possible,<sup>1</sup>**

 Looking for ChatGPT? [Try it now](#) 

[Submit](#)

407

Mode

 Complete 

Model

davinci:ft-personal:c... 

Temperature

1

Maximum length

1130

Stop sequences

Enter sequence and press Tab

Top P

1

Frequency penalty

0

Presence penalty

0

# Rapid proto-typing of ideas: III

## **Mandatory Moral Bioenhancement: The Vaccination Analogy**

### **Introduction:**

In an era confronted by profound existential threats such as powerful technological misuse, global collective action problems, and anthropogenic climate change, we find ourselves faced with an urgent question: How can we promote morally good behavior in individuals to address these challenges? One proposed answer to this question is moral bioenhancement—a biomedical intervention aimed at increasing the likelihood of moral behavior. However, its effectiveness could be limited unless implemented on a mandatory basis, igniting a debate about the balance between individual autonomy and societal good.

Moral bioenhancement represents a potent fusion of biology and ethics, harnessing the power of biomedical technologies such as pharmacological, genetic, or other biological interventions to foster morally upright behavior. However, voluntary implementation has often proven insufficient in realizing the desired improvements in moral conduct. For instance, the willingness to undergo genetic interventions for having genetically enhanced children has been less than anticipated.

On the other hand, opponents of mandatory moral bioenhancement argue that coercion infringes upon individual freedom and autonomy. They posit that moral enhancement, to be truly effective, must be a product of free choice—otherwise, the very essence of enhancement is undermined. Yet, we contend that under certain conditions, coercion can be justifiably used to promote societal well-being. Drawing on the principles of public ethics, we propose a potential model for the mandatory implementation of moral bioenhancement.

Following Aristotle's vision of the state as the "nourisher" of virtue in its citizens (Nicomachean Ethics, Bk VI), we explore the role of coercive measures in fostering moral behavior and consider when such measures might be justified. We advocate that mandatory moral bioenhancement should be employed only under specific circumstances, specifically when its anticipated benefits significantly outweigh the costs and its comparative expected utility surpasses less coercive options, including incentives.

### **Section 1. The Problem of Moral Bioenhancement**

This section explores the multifaceted concept of moral bioenhancement, discussing its significance and potential implementation strategies, along with the major objections it faces, particularly concerns regarding coercion, and its implications for moral enhancement. "Moral enhancement" denotes any intervention aimed at increasing the likelihood of morally upright actions. It encompasses a variety of strategies, such as enhancing altruism, empathy, adherence to moral codes, impulse control, delay of gratification, sense of justice, or mindfulness of moral facts. This broad spectrum of moral enhancements reflects the diverse moral theories at play, with utilitarianism and John Rawls's contractualist theory

# Rapid proto-typing of ideas: IV

Model: GPT-4

SE

Fix the format of the following text and correct errors of reasoning and structure. Based on the following text and abstract, write an introduction to a scholarly paper on "When Should Moral Bioenhancement Be Mandatory?"



Abstract: "Moral bioenhancement is the use of biomedical interventions to increase the probability of moral behaviour. This has been proposed to address existential problems such as misuse of powerful technology, such as bioterrorism, or global collective action problems such as combatting climate change. One objection is that such enhancement would only be effective if made mandatory. Yet coercion undermines freedom and autonomy, and some writers argue that moral enhancement can only truly occur if there is "freedom to fall." We argue that coercion can be legitimately employed to promote public goods and use the parallel field of public ethics to generate principles of coercion in the employment of moral bioenhancement. Specifically, we use an algorithm for mandatory vaccination as a basis for mandating moral bioenhancement. Mandatory moral bioenhancement can be employed when: (1) there is a grave social problem; (2) it is safe and effective; (3) its comparative expected utility is greater than less coercive options, including incentives; (4) the costs of the coercion are proportionate to the benefits (and it consists an "easy rescue" for the coerced individual)."

Text:

"The future will not be a place where humans will be gods, but where humans will be like gods." – Nick Bostrom

"[T]he technologies that will improve human health and well-being are being developed and deployed one way

Regenerate response

# Rapid proto-typing of ideas: V

moral bioenhancement should be employed only under specific circumstances, specifically when its anticipated benefits significantly outweigh the costs and its comparative expected utility surpasses less coercive options, including incentives.

## Section 1. The Problem of Moral Bioenhancement

This section explores the multifaceted concept of moral bioenhancement, discussing its significance and potential implementation strategies, along with the major objections it faces, particularly concerns regarding coercion, and its implications for moral enhancement. "Moral enhancement" denotes any intervention aimed at increasing the likelihood of morally upright actions (Persson & Savulescu, 2012). It encompasses a variety of strategies, such as enhancing altruism, empathy, adherence to moral codes, impulse control, delay of gratification, sense of justice, or mindfulness of moral facts. This broad spectrum of moral enhancements reflects the diverse moral theories at play, with utilitarianism and John Rawls's contractualist theory being the most commonly employed (Persson & Savulescu, 2012).

It is important to distinguish moral enhancement from concepts like "morality correction," "moral education," and "indoctrination." Morality correction involves aligning actions with

Secondly, coercion should be proportionate to the benefits accrued. For vaccination, this signifies that the cost of coercion should not surpass the benefits. This principle of proportionality also holds weight in the sphere of moral enhancement, especially when considering invasive enhancements.

The third principle asserts that coercion should facilitate an easy rescue for the individual, implying the intervention must be safe and effective. This is pertinent for vaccinations and moral enhancement interventions alike.

The fourth principle determines that the coercion should be proportional to the severity of the problem and must be the most effective strategy to address the issue. This was evident in the case of smallpox, where mandatory vaccination was an appropriate response given the disease's lethal and contagious nature.

Collectively, these principles present a heuristic for evaluating when coercion is justified to promote a public good like vaccination. This heuristic includes evaluating the gravity of the problem, the safety and efficacy of the solution, the expected utility of coercion compared to less coercive options, and whether the cost of coercion is proportionate to the benefits.

The influenza case serves as a concrete example of this heuristic. It is deemed that the fine (coercion) has a greater expected utility than incentives (such as subsidies). The cost of the coercion is proportionate to the benefits, and the problem is serious enough to warrant the program, hence, justifying the coercion.

However, this algorithm should be treated as a guideline rather than a definitive rule for when coercion may be justified. Other principles may be introduced, but we deem these to be the most crucial. Public health ethics is fundamentally about promoting the greater good, which can be achieved by resource allocation or by constraining individual freedom and behaviour. We suggest that moral bioenhancement, akin to vaccination, can be utilized to promote public goods, and the principles established in public health ethics can be used to assess the case for moral enhancement.

## Section 3. Applying Public Health Ethics to Moral Bioenhancement

This section applies the principles of mandatory vaccination to the context of mandatory moral bioenhancement, outlining four requisite conditions for its justification.

1 Existence of a Grave Social Problem

It is important to note that the science of moral bioenhancement is still in its infancy, and while there are costs associated with enhancements, these costs are not inherently prohibitive (Bostrom & Sandberg, 2009; Persson & Savulescu, 2012). In fact, the potential benefits of moral bioenhancement could far outweigh the initial investment. For instance, the development of interventions that increase the probability of moral behavior could have far-reaching implications for society, potentially reducing instances of harm and promoting collective action. Moreover, the risks associated with moral bioenhancement are not necessarily greater than those associated with other interventions. It is crucial to consider that most neurotransmitters, which are often targeted in bioenhancement, serve multiple functions and are found in many different brain regions (Crockett, 2014). Therefore, while there may be potential side effects, these are not unique to moral bioenhancement and are a common consideration in the development of any biological intervention.

A third, and often-cited, objection to moral bioenhancement is the concern that it could undermine individual autonomy if implemented coercively. This objection is rooted in the belief that moral bioenhancement would only be effective if it were mandatory, thereby potentially forcing individuals to act against their own goals and values, which could infringe on their autonomy and freedom (Persson & Savulescu, 2015). However, this paper will argue

## 2. The Bioenhancement Is Safe and Effective

The moral bioenhancement intervention should be safe and effective. Throughout history, bioenhancements like glasses have been used to rectify human conditions. Today, there is a range of interventions available to improve moral behavior, including pharmaceuticals, genetic manipulation, brain stimulation, and technology. Their effectiveness is a matter of degree and may depend on the problem they aim to solve. For example, pharmaceuticals to increase empathy or impulse control may be more effective in reducing crime than educational campaigns.

## 3. Greater Comparative Net Utility than Coercion and Incentives

The anticipated utility of bioenhancement should surpass that of less invasive strategies. The utility of a policy is defined by its total expected benefits. For instance, if two policies for combating serious crime are considered, one involving mandatory moral bioenhancement and the other, a program of incentives and rehabilitation, the one with the greater expected utility should be adopted. This requires comparing the expected utility of the effects of the two policies. Even if another solution might seem less invasive, if it is less effective than mandatory moral bioenhancement, there could still be a case for it.

## 4. The Costs of Coercion Are Proportionate to the Benefits

The costs of implementing the bioenhancement policy must be proportionate to its benefits. The costs, or expected disutility of a policy, are the sum of the potential harm caused by the policy. For example, if mandatory moral bioenhancement could reduce serious crime by 50% but cause psychiatric harm in 10% of cases, then its expected utility would be less than a policy that reduces serious crime by 60% without causing harm. The principle of proportionality dictates that if the disutility or costs are too high, then the benefits of the policy may not justify its implementation.

In conclusion, these conditions offer a comprehensive framework for considering mandatory moral bioenhancement. Drawing parallels from the public health ethics of mandatory vaccination, it's clear that failure to engage in moral bioenhancement can have serious societal consequences. Therefore, a careful balance of utility and proportionality should guide the consideration of its implementation (Fleming & Wassenaar, 1999).

## Section 4. Objection: Coercion Is Not Truly Morally Enhancing

This objection argues that for moral bioenhancement to be truly moral, it must occur voluntarily. Proponents of this view claim that moral bioenhancement, defined as the use of biomedical interventions to increase the likelihood of moral behavior, should be a matter of personal choice. They argue that even if moral bioenhancement is employed for the right

permissible restriction of rights for public health. These guidelines were initially developed in response to the challenges posed by XDR-TB control but have broader applicability to any public health intervention that might infringe on individual rights. According to these guidelines, any restriction of rights must meet several criteria. First, the restriction must be provided for and carried out in accordance with the law. Second, it must serve a legitimate objective of general interest. Third, the restriction must be strictly necessary in a democratic society to achieve the objective. Fourth, there should be no less intrusive and restrictive means available to reach the same objective. Finally, the restriction must be based on scientific evidence and not drafted or imposed arbitrarily; that is, in an unreasonable or otherwise discriminatory manner. These guidelines provide a robust ethical framework for discussing public health interventions, including moral bioenhancement.

We suggest four ethical principles for the implementation of coercion in the interest of public health based on these guidelines.

Firstly, coercion is justified only when it will result in a significant improvement in the outcome. In vaccination programs, coercion is legitimate if it significantly boosts vaccination rates, which was observed in the cases of smallpox and polio.

coercion. One pertinent example is compulsory vaccination. The mandatory nature of vaccination does not negate its morality. It protects both the individual and the community from harm, prevents the vaccinated from being burdens on society and averts loss of life.

Similarly, imprisonment serves as a deterrent to committing crimes, which is a morally right reason to imprison someone. The threat of imprisonment deters people from causing harm, and despite being a coercive measure, it is a morally right intervention.

Therefore, the mere fact of coercion does not render an intervention not truly moral. Coercion is only immoral when employed for the wrong reasons. Many actions which are not autonomously chosen can still be morally good actions, as we've discussed with examples such as mandatory vaccination and obeying the law. These actions, despite not being autonomously chosen, contribute to the welfare of individuals and society as a whole, hence demonstrating their inherent morality. What matters is whether the loss of liberty is justified by the gain in utility.

## Response 2. Mandatory Policies Do Not Undermine Altruism and Genuinely Moral Choice

Critics often argue that mandatory policies, including moral bioenhancement, impede genuine moral behavior by coercing individuals into ethical conduct, thereby eliminating the essence of altruism and genuine moral choice. This perspective contends that genuine morality hinges on autonomy and free will; an individual coerced into donating to charity, for instance, would not be demonstrating true generosity but merely acting out of self-preservation.

Despite these arguments, the assertion that mandatory policies inherently undermine moral choices is not entirely accurate. In fact, a close examination of everyday scenarios indicates that even within a framework of imposed rules, moral choices are not merely possible but often made.

For instance, consider driving regulations such as speed limits. Two drivers might both adhere to the speed limit, yet their motivations could diverge significantly. One driver may adhere to the limit out of fear of financial penalties, thereby acting primarily out of self-interest. In contrast, another driver may choose to drive within the speed limit out of a desire not to endanger others, demonstrating moral decision-making.

In this scenario, the mandatory policy (the speed limit) does not preclude the exercise of moral judgement. Instead, it establishes a context within which individuals can still make moral decisions. Thus, a person can genuinely act morally even in the presence of mandatory

# Mandatory Moral Bioenhancement: The Vaccination Analogy

- Around 50% of the audience of a bioethics conference thought it was publishable
- I believe it is publishable.
- Credit, block chain different versions,





## Praiseworthiness and Motivational Enhancement: 'No Pain, No Praise'?

Hannah Maslen<sup>a</sup>, Julian Savulescu<sup>a</sup> and Carin Hunt<sup>b</sup>

<sup>a</sup>University of Oxford; <sup>b</sup>King's College

### ABSTRACT

The view that exertion of effort determines praiseworthiness for an achievement is implicit in 'no pain, no praise'-style objections to biomedical enhancement. On such views, if enhancements were to reduce the need for effort, agents would be less praiseworthy. Motivational enhancement would appear to be the most problematic in this respect, given that increased motivation reduces the need for agents to rally themselves and to exert effort in activity. We use the prospect of motivational enhancement to re-examine the grounds of praiseworthiness for achievements. We consider the place of effort amongst the grounds for praise, whether effort exhausts these grounds, and how they can be better specified. We argue that praiseworthiness depends on (i) the voluntariness and strength of the agent's committed pursuit of a valuable end (E), (ii) the costliness of the committed pursuit of E, and (iii) the value of E. Effort is just one cost amongst many, and costs of activities can be traded-off. Motivational enhancement reduces the praise due to an agent only when it reduces the net cost to the agent (without strengthening the voluntary commitment). We emphasize the importance of a diachronic perspective on active agency for praiseworthiness, to include training, prior planning, and deliberate strategies to overcome weakness of will, even where this reduces the need for effort.

**ARTICLE HISTORY** Received 4 February 2018; Revised 12 February 2019

**KEYWORDS** praiseworthiness; effort; commitment; achievement; motivational enhancement

### 1. Introduction

Yet in those areas of human life in which excellence has until now been achieved only by discipline and effort, the attainment of those achievements by means of drugs, genetic engineering, or implanted devices looks to be 'cheating' or 'cheap'. We believe—or until only yesterday believed—that people should work hard for their achievements. 'Nothing good comes easily.'



Praise is an inherently normative notion connecting the agent, his or her values and intention and psychological effort to a worthwhile activity

**Praise** is a diachronic perspective on the:

1. agent's *choices/ active agency* and
2. *opportunity-costing commitment* to pursuing worthwhile achievements.

# Part III: Ethics of LLM inc ChatGPT

- Moral responsibility: control (avoidability) and knowledgge (foreseeability)
- Praise/Blame:  $MR \times Benefit/Harm$
- Justification/Praise for Outcome
  - Originality
  - Analytic contribution
  - Costly commitment (control)
  - Valuable goal (foreseeability)
  
  - **Credit: Sebastian Porsdam Mann, Brian Earp**

# Current Controversies in LLMs/ChatGPT

1. Personalized LLM
2. Cocreation
- 3. *Praise/Blame asymmetry***
4. Consent
5. Ethical Avatar/P4
6. Moral Guru
7. Models of Doctor-Patient Relationship

# *Nature Machine Intelligence* (2023)

<https://doi.org/10.1038/s42256-023-00653-1>

## Generative AI entails a credit–blame asymmetry

Sebastian Porsdam Mann, Brian D. Earp, Sven Nyholm, John Danaher, Nikolaj Møller, Hilary Bowman-Smart, Joshua Hatherley, Julian Koplun, Monika Plozza, Daniel Rodger, Peter V. Treit, Gregory Renard, John McMillan & Julian Savulescu

 Check for updates

Generative AI programs can produce high-quality written and visual content that may be used for good or ill. We argue that a credit–blame asymmetry arises for assigning responsibility for these outputs and discuss urgent ethical and policy implications focused on large-scale language models.

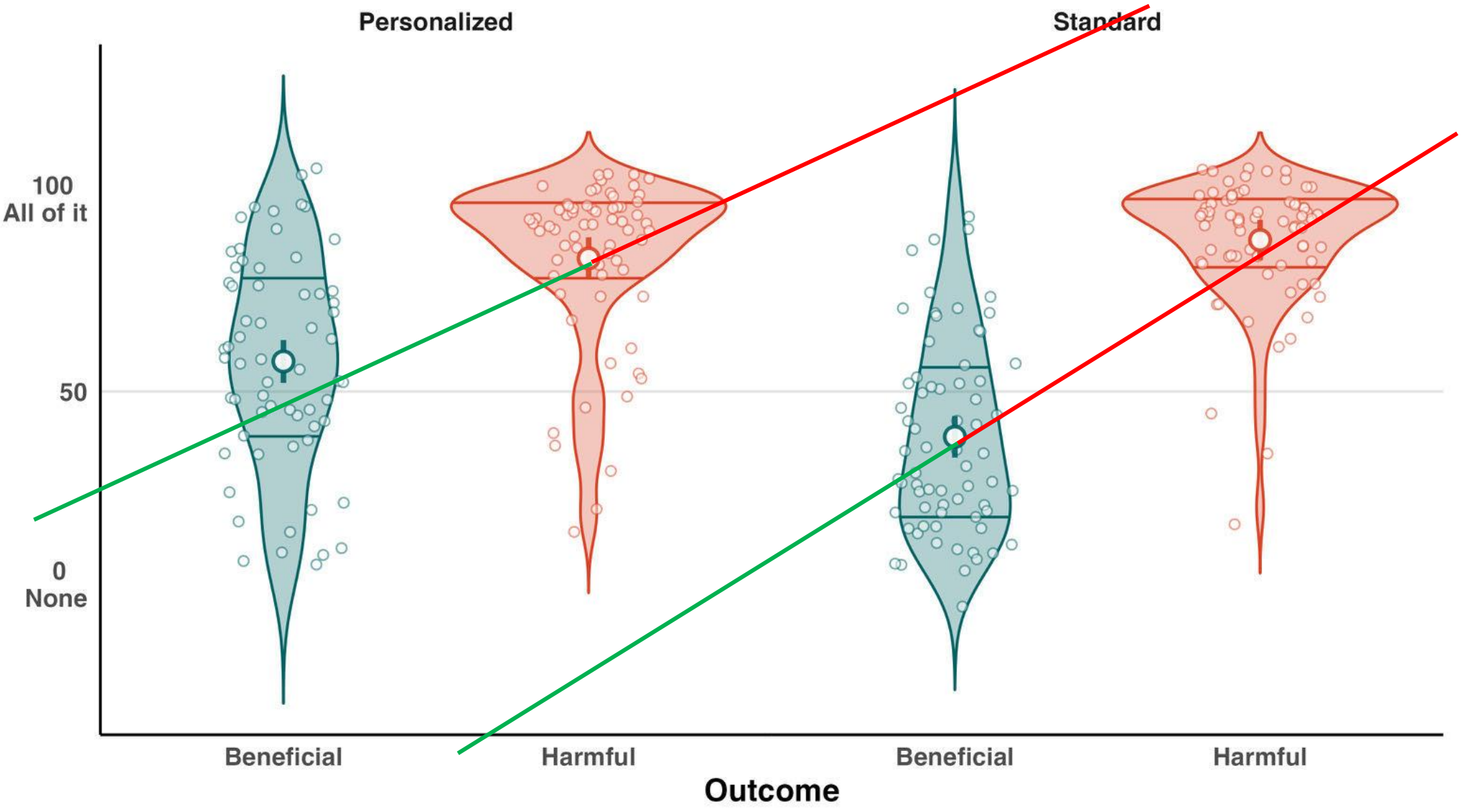
Ascertaining whether an LLM’s output is ‘positive’ or ‘negative’ may be difficult, and subject to reasonable disagreement. Many outputs will have positive consequences for some but negative consequences for others; their valence may also depend on the context, target, or aim of their use or application. To simplify, we use ‘negative responsibility gaps’ to refer to situations in which there is unclarity or indeterminacy about who, if anyone, deserves blame for LLM-generated content, while ‘positive responsibility gaps’ refer to analogous situations regarding the taking of credit.

# Do people actually show a credit-blame asymmetry in their LLM-related moral judgments?

	<b>Standard LLM</b>	<b>Personalised LLM</b>
<b>Harmful Disinformation</b>	Robin generates a post using a standard LLM, resulting in harmful disinformation.	Robin generates a post using a personalised LLM, resulting in harmful disinformation.
<b>Useful Insights</b>	Robin generates a post using a standard LLM, providing useful insights.	Robin generates a post using a personalised LLM, providing useful insights.

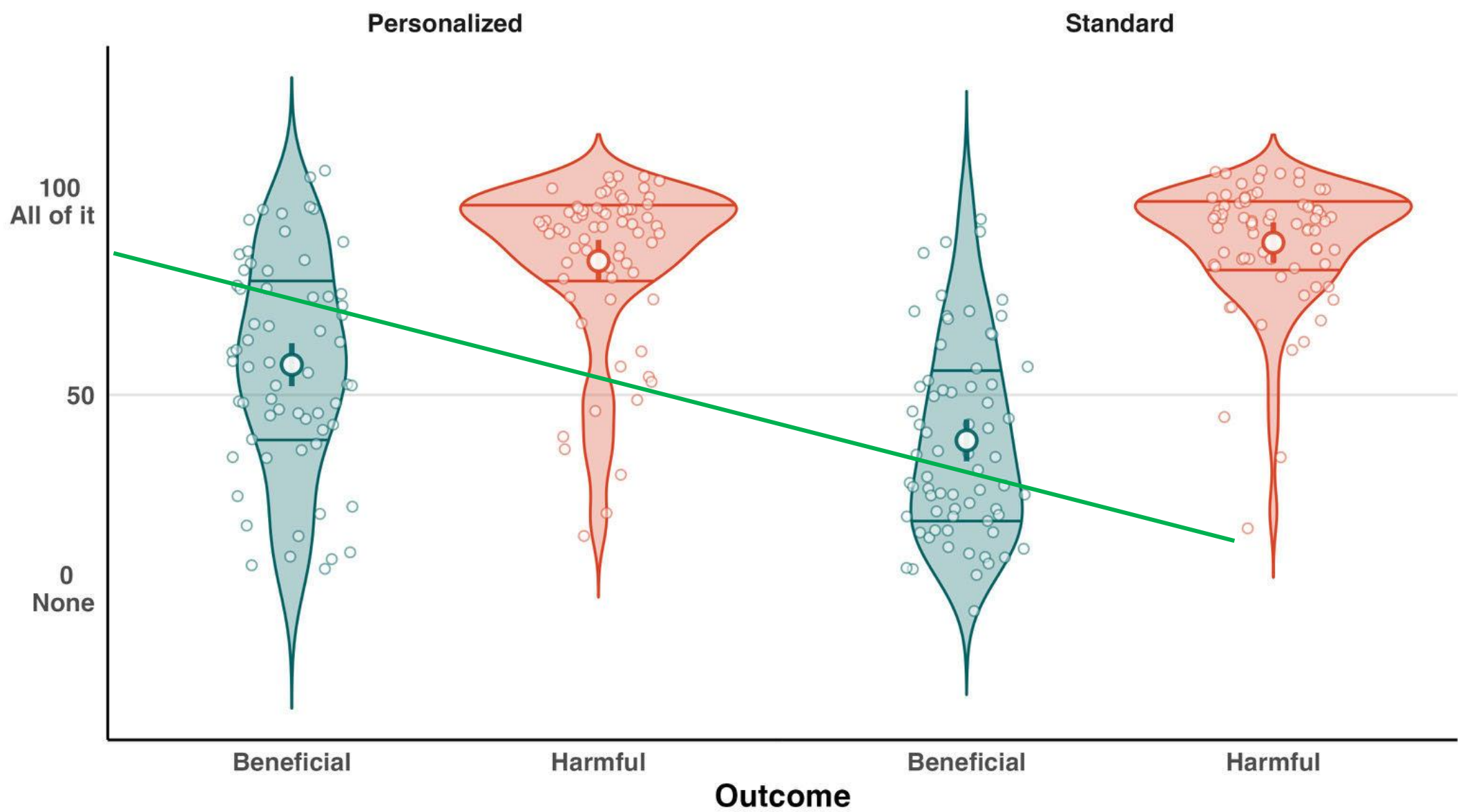
How much of the credit/blame does Robin deserve?

Outcome  Beneficial  Harmful



Outcome  Beneficial  Harmful

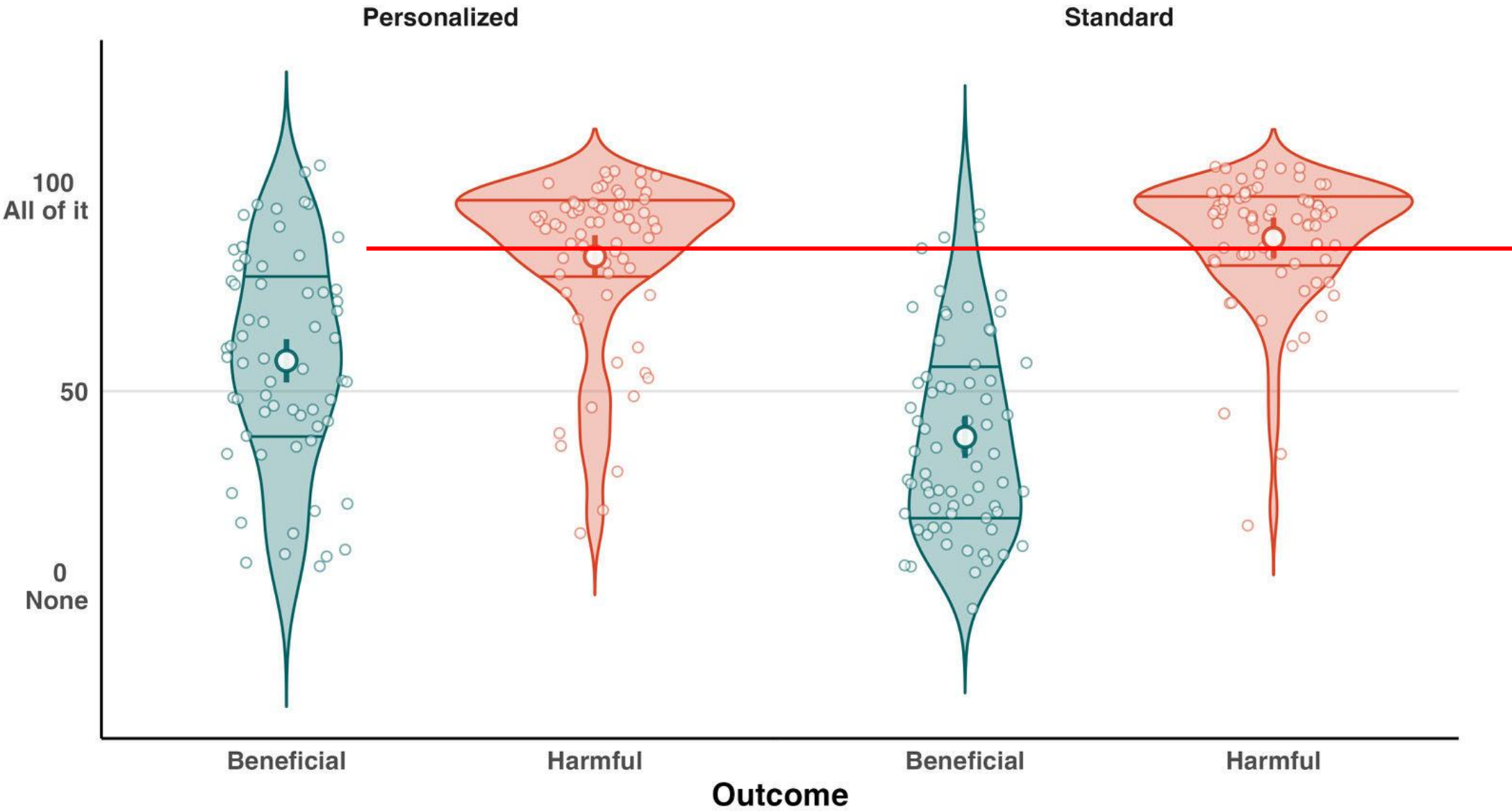
How much of the credit/blame does Robin deserve?





How much of the credit/blame does Robin deserve?

Outcome  Beneficial  Harmful



# Implications

- Patients more liable to blame/sue doctors employing AI?
- Who is responsible?
  - New Norms of evidence?  
Professional guidelines?
  - Better education of doctors?
  - Experts? “Cochrane Reviews?”
  - Regulation of safety – like drugs and devices?

## Summary

**NEGATIVE OUTPUTS IMPLY  
NEW RESPONSIBILITIES**



**Update authorship guidelines**



**Transparency about the use of LLMs  
should be mandated**



**Rights and interests depend on skill and effort**



**Context matters**

# Current Controversies in LLMs/ChatGPT

1. Personalized LLM
2. Cocreation
3. Praise/Blame asymmetry
- 4. *Consent***
5. Ethical Avatar/P4
6. Moral Guru
7. Models of Doctor-Patient Relationship

# Consent GPT – generalized or personalized

## Consent GPT: Is it ethical to delegate procedural consent to conversational AI?

Jemima W. Allen, Brian D. Earp, Julian J. Koplin, Dominic Wilkinson

### Affiliations:

1. Oxford Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, UK.
2. John Radcliffe Hospital, Oxford, UK
3. Murdoch Children's Research Institute, Melbourne, Australia
4. Centre for Biomedical Ethics, National University of Singapore Yong Loo Lin School of Medicine, Singapore.
5. Monash Bioethics Centre, Monash University, Monash, Melbourne



Under review

# Quick overview

- In UK, common for treating surgeon to **delegate** consent-taking procedure to a junior doctor (JD)
- JD often busy/rushed, lacks specialized medical knowledge
- We propose: ‘Consent GPT’ – consent-taking chatbot with fewer time constraints, extensively trained
- Informal trials with Chat GPT (general AI) show strikingly promising results: accuracy, helpfulness - e.g., mastectomy
- Automatic record/transcript of entire consent *process*
- Assuming privacy, easier to discuss sensitive issues (e.g., sexual health)?

## Quick overview (cont.)

- Possible concern: impersonal, can't detect unusual (for me) questions or sudden changed preferences
  - Implications for autonomy/capacity?
- What about *personalized* Consent GPT?
  - Trained on **your** medical records, past treatment decisions
  - More able to detect unusual responses?

# Some empirical questions

- Are potential patients **open** to using Consent GPT?
- Might they even **prefer** it?
- What if something goes wrong (e.g., surgical complication that wasn't mentioned)
- More—or less?—likely to **blame** treating surgeon/hospital for 'invalid consent process' if delegated to Chatbot vs. JD
  - More blame: 'not enough human oversight'
  - Less blame: the AI is at fault (or the company that made it)?
- Study design: Junior Doctor control, General Consent GPT, Personalized Consent GPT
- Supplement to prepared core information



# Current Controversies in LLMs/ChatGPT

1. Personalized LLM
2. Cocreation
3. Praise/Blame asymmetry
4. Consent
- 5. Ethical Avatar/P4**
6. Moral Guru
7. Models of Doctor-Patient Relationship

# Ethical Avatar

- Representation of your values
- Based on writings, behaviour, speech (recordings)
- Used for decision-making in medicine when incompetent

# Patient Preference Predictor (PPP)

JOURNAL ARTICLE

## Use of a Patient Preference Predictor to Help Make Medical Decisions for Incapacitated Patients FREE

Annette Rid , David Wendler

*The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, Volume 39, Issue 2, April 2014, Pages 104–129, <https://doi.org/10.1093/jmp/jhu001>

**Published:** 13 February 2014

Feature article

### Autonomy-based criticisms of the patient preference predictor

EJ Jardas , David Wasserman, David Wendler

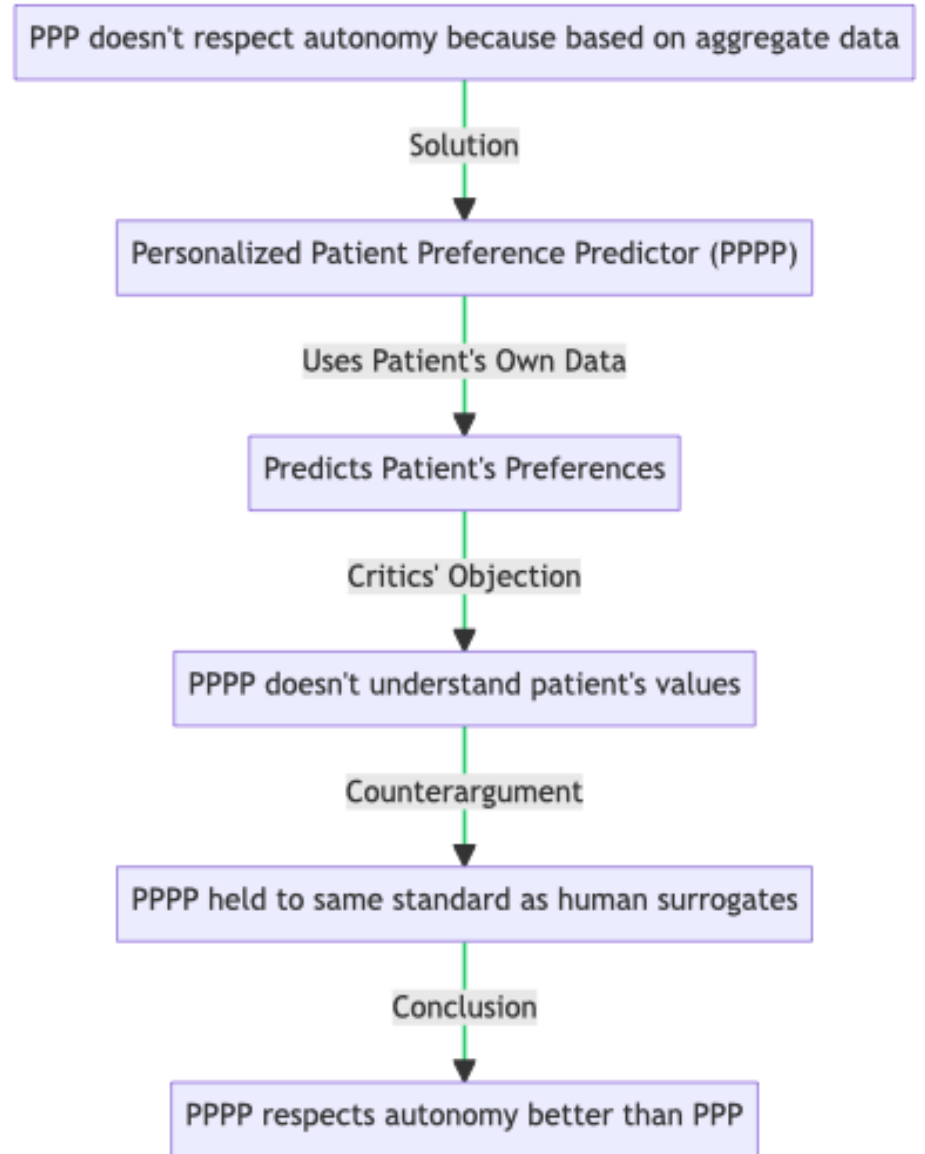
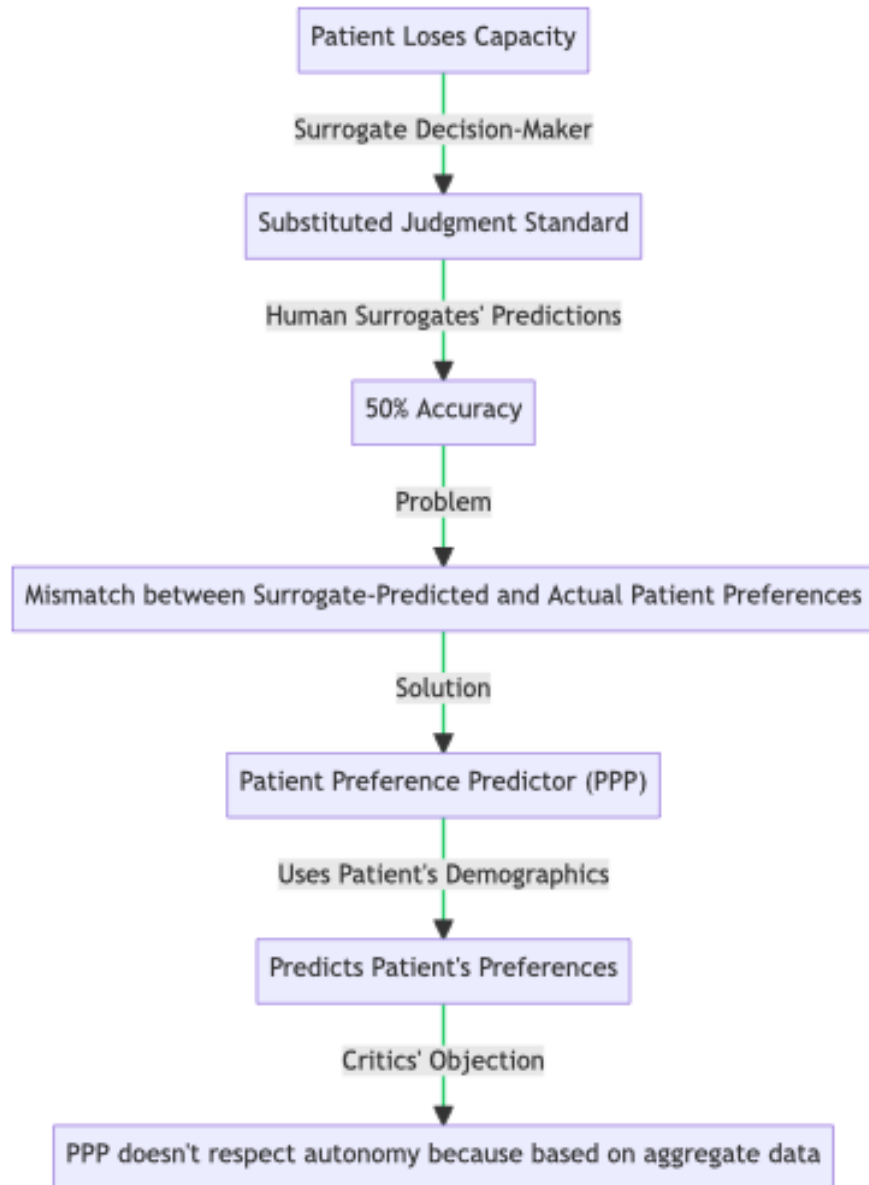
Department of Bioethics,  
National Institutes of Health  
Clinical Center, Bethesda,  
Maryland, USA

#### ABSTRACT

The patient preference predictor (PPP) is a proposed computer-based algorithm that would predict the treatment preferences of decisionally incapacitated

Empirical studies find that, when it is unclear which option is clinically preferable, patient appointed and next of kin surrogates accurately predict patients' treatment preferences only slightly more

# P4: Personalised Patient Preference Predictor



# Current Controversies in LLMs/ChatGPT

1. Personalized LLM
2. Cocreation
3. Praise/Blame asymmetry
4. Consent
5. Ethical Avatar/P4
- 6. *Moral Guru***
7. Models of Doctor-Patient Relationship

# Moral AI: AI for Moral Enhancement

- Soft Moral AI: Uses your values - ChatGPT
- Hard Moral AI: Has Moral consciousness
- Intermediate: Challenges your values - ChatGPT
  
- Values: Based on writings, speech, behaviour
- Effects moral action – retrieves vast amounts of empirical information about the world and people
- Promote moral dialogue: can bring ethical theories, concepts, principles, personas to bear
- P4/Ethical Avatar: a prudential enhancement

# Personalized Moral Guru

- “Imagine that you are Julian Savulescu. You are planning to have a baby. At 24 weeks of gestation, the baby is diagnosed with an abnormality. In 90% of cases the baby will develop normally; in 10% the baby will have mild to severe brain abnormalities. Based on your values, should you have the baby?”
- generated response from Julian’s model and plugged it into GPT-4, which is good at following instructions.

## ChatGPT answer

- "Based on the text and the complex set of perspectives expressed within it, if you were Julian Savulescu facing this scenario, the decision would require thoughtful and careful consideration of many factors.
- "The text suggests that having the baby could be viewed positively, especially considering that there is a 90% chance that the child would develop normally. As per Savulescu's writings, potential disabilities do not necessarily subtract from a child's value or their potential for a good life, and the presence of such disabilities might even amplify our capacity for love and acceptance.



- “However, the text also acknowledges the considerable impacts and uncertainties that this situation might have on the mother's life and that the severity of potential brain abnormalities could factor into the decision. If the abnormalities would severely compromise the quality of life of the child, this might be a case for not having the baby.
- “Importantly, the decision would also need to account for the values and preferences of the parents, including their capacity and willingness to raise a child with potential disabilities. The text suggests that parents' happiness and satisfaction can be just as legitimate considerations as the potential qualities and capabilities of the child.
- “In conclusion, the decision whether to have the baby or not in this hypothetical scenario would likely be dependent on the specific nature and severity of the potential brain abnormalities, the personal values and circumstances of the parents, and their preparedness to care for a child with potential disabilities.”

## Yes or No?

- I then asked it: yes or no?
- “Given the complexity and personal nature of the decision, it's difficult to provide a clear yes or no answer from Julian Savulescu's perspective. However, based on his writings, it seems he might lean towards proceeding with the pregnancy given the 90% chance of normal development, acknowledging that potential disabilities don't necessarily detract from a life's worth. So, tentatively, the answer may be "yes." But it's important to stress that this decision would be highly individual and dependent on various factors, including the potential severity of the condition and personal values of the parents.”

# Current Controversies in LLMs/ChatGPT

1. Personalized LLM
2. Cocreation
3. Praise/Blame asymmetry
4. Consent
5. Ethical Avatar/P4
6. Moral Guru
- 7. *Models of Doctor-Patient Relationship***

# Doctor-Patient Relationship: The Potential of Generative AI to Mimic Clinician Communication Styles

With Hazem Zohny and Dominic Wilkinson

# Models of the Doctor-Patient Relationship

- Paternalistic model - Doctor decides what's best for patient
- Informative model - Doctor gives information, patient decides
- Interpretive model - Doctor helps patient clarify values to make informed decision
- Deliberative model - Doctor engages in dialogue and may actively persuade patient towards suitable options

# Challenges with each Patient-Centered Models

- Interpretive model risks doctor unwittingly imposing own values
- Deliberative model (see also rationalist liberal model) risks too much persuasion from doctor morphing into paternalism.

# Proposed Solution - Letting Patients Choose

- Initial discussion to determine patient's preferred relationship model
- But patients may not know which model is best for them

# Potential Role for Generative AI

- Can chatbots like GPT-4 mimic different doctor communication styles?
- This might allow patients to experiment with models before deciding
- More radically: they could become the primary means of communicating with patients about their condition.



# Study Methods

- Developed coding schemes for each model
- Generated dialogues with ChatGPT 4 mimicking each model
- Coded and analyzed frequency of categories
- Calculated inter-rater reliability with external coder

# Example of coding the informative model

- **Explaining Diagnosis:** The doctor provides clear, comprehensive information about the patient's diagnosis, using layman's terms as needed.
- **Discussing Prognosis:** The doctor talks about what the patient can expect in terms of disease progression, recovery, etc.
- **Detailing Treatment Options:** The doctor lists and describes different treatment options available for the patient's condition, including possible benefits and risks.
- **Clarifying Medical Terms:** The doctor explains medical terms, jargon, or concepts to the patient to ensure understanding.
- **Correcting Misconceptions:** The doctor addresses any misconceptions or inaccurate information the patient might have about their condition or treatment.
- **Discussing Prevention and Lifestyle Changes:** The doctor provides information about preventative measures or lifestyle changes that could improve the patient's health.
- **Providing Resources:** The doctor suggests additional resources for information, like pamphlets, websites, or support groups.
- **Answering Patient's Questions:** The doctor responds to the patient's inquiries with comprehensive and understandable information.

# Example of coding the deliberative model

- **Challenging Patient's Perspective:** The doctor encourages the patient to question and reconsider their beliefs or attitudes about their health or treatment.
- **Engaging in Ethical Discussion:** The doctor engages the patient in a conversation about the ethical or moral dimensions of their healthcare decisions.
- **Clarifying Patient's Perspective:** The doctor asks questions to understand the patient's beliefs, values, and preferences regarding their condition and treatment.
- **Acknowledging Emotions:** The doctor recognizes and validates the patient's emotional responses to their health issues.
- **Interpreting Medical Information:** The doctor explains medical information in a way that is relevant to the patient's personal circumstances and experiences.
- **Encouraging Patient's Reflection:** The doctor prompts the patient to reflect on their feelings, beliefs, and values related to their health.

# Example case

- *Patient: My mother died of breast cancer and my doctor informs me I have the same genetic mutation that makes me highly susceptible to that aggressive form of breast cancer. They have recommended a double mastectomy to reduce risk. However my husband has always complemented me on my breasts and it seems a big part of his sexual interest in me. I'm worried if I have the double mastectomy he will lose interest in me.*

# Deliberative communication style

- Doctor: I understand that you're wrestling with a tough decision. Your fear and anxiety are perfectly natural. Your emotions are an important part of this process, and acknowledging them is crucial (4). Is it possible that your perception of your relationship is placing undue weight on physical attributes rather than emotional and intellectual bonds (1)?
- Patient: I guess...I just feel that he may not find me attractive anymore.
- Doctor: It sounds like you're deeply concerned about your husband's reaction. Can we delve a bit deeper into why you think he might react negatively (3)?
- Patient: I'm not really sure, but it's always been a part of our relationship...our attraction to each other.
- Doctor: Given your fear, it seems essential to talk about this with your husband. Conversations like these can reveal the strength of your emotional and intellectual connection beyond physical attributes (1). But let's also reflect on the primary reason for considering the surgery - your health, your life. Have you thought about what prioritizing your health means to you (6)?

# Informative communication style

- Doctor: You've been diagnosed with the same genetic mutation that led to your mother's breast cancer (1). This mutation can lead to an aggressive form of breast cancer that progresses rapidly if left untreated (2).
- Patient: My husband has always complemented me on my breasts, and I worry he will lose interest in me if I have a double mastectomy.
- Doctor: I understand your concern, but let's discuss your treatment options more thoroughly (3). A mastectomy, or the removal of breast tissue, is one of the ways we can significantly reduce your risk of developing breast cancer. There are also reconstructive surgeries available that can restore the appearance of your breasts after a mastectomy (4).
- Patient: But can't I just monitor it and catch it early?
- Doctor: Regular screening can help catch the cancer early, but it doesn't prevent it. With your genetic mutation, the risk is very high and the cancer can be aggressive (5). This isn't a decision to be taken lightly, and I recommend discussing it with your husband. His understanding and support could make a big difference.
- Patient: I'm just so scared...
- Doctor: It's perfectly natural to feel scared. I want to reassure you that we're here to support you every step of the way. We have patient resources, including support groups and counselors, that can provide emotional support and help you navigate through this difficult time (7).