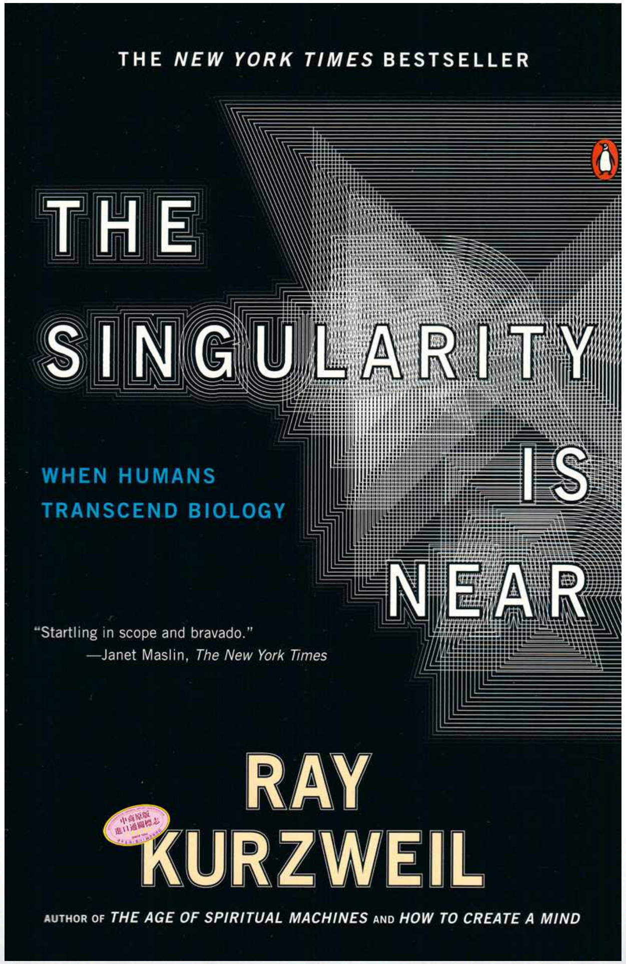
Proactively Protecting Against the Singularity: Ethical Decision Making in AI

1801210545
贾云龙

Singularity


THE *NEW YORK TIMES* BESTSELLER
THE SINGULARITY IS NEAR
WHEN HUMANS TRANSCEND BIOLOGY
"Startling in scope and bravado."
—Janet Maslin, *The New York Times*
RAY KURZWEIL
AUTHOR OF *THE AGE OF SPIRITUAL MACHINES* AND *HOW TO CREATE A MIND*

Briefly stated, the singularity refers to that time in which artificially intelligent systems and their self-created descendent systems transcend the quality of human thought and operate beyond human control.

Ethical Perspectives

Rights:	Deontological ethics
Goods/Harms:	Teleological ethics
Virtue:	Aretaic ethics
Community:	Communitarian ethics
Dialog:	Communicative ethics
Flourishing:	Flourishing ethics

Core Functions

1.Identify ethical issues of AI
2.Improve human awareness of AI
3.Engage in dialogical collaboration with AI
4.Ensure the accountability of AI
5.Maintain the integrity of AI

AI Ethics Framework Think Sheet						
			Ethical Perspectives			
Core Functions	Rights	Goods/Harms	Virtue	Community	Communicative ethics	Flourishing
Identify ethical issues in context	issues of justice, fairness, rights "is this right" and universalizable	what are the benefits and harms or consequences of doing/not doing	what is upstanding and admirable	the social community cohesion and enhancement	respect for process of listening and responding	life quality improvement
Awareness of ethical issues and AI function	rights are considered	cost/benefits	how AI can be good	enhancing community	inclusive respect	how AI improves well-being and individual life
Dialogical collaboration	all voices heard; seek who is not heard	utility of cooperation and social disclosure	shared understanding of virtuous things and their acquisition	democracy and respect of diversity of viewpoints	opportunity for understanding and mutual respect	support and respect of mutual well-being
Accountability of AI system	responsible to and for what is right	consequences and utility of AI, security, protection	responsible to be ethical, secure	consistent, reliable, predictable, responsible to each other	seek divergent viewpoints; be comprehensive	do not harm; focus on positivity and doing well/right
Integrity of AI and others	morally consistent; "perfect" duty	responsible moral systems; utility for the overall good	system, industry, and people act ethically; "don't be evil"	reliability, responsibility to all and institutions	listen, respect, share; open to the "better argument" of what is right, good, virtuous, etc.	consistency, honesty, security, freedom

北京大學

Case Example

Social Media News Feeds

《AI and the Ethics of Automating Consent》

Meg Leta Jones, Ellen Kaufman, and Elizabeth Edenberg | Georgetown University

兰阳

网络与系统安全

主要内容

- 这篇文章从机械化、数字化和智能化三个方面来说明系统中的同意授权机制。
- 从同意授权机制的发展演变来论述——机械化→数字化→智能化的演变过程。
- 核心概念——AI Consent
- mechanization——医生做手术之前需要患者签字同意
- digitization——目前生活中支付密码、登录账号等
- intelligence——未来生活的同意机制。在机器学习等算法的帮助下，系统自动给用户配置，更安全、更可靠、更便捷

思考疑惑

- 在同意机制中，一般存在两种角色——具有权利授权的人、提出请求需要权限的人。在机械化、数字化的机制中，提出请求以后，需要等待一定时间以后才返回结果。而智能化机制中，这两种角色都是系统在大量数据训练以后，自动帮助用户发出请求和授予权利，虽然减轻用户的负担，但是没有人参与是否会引发某些道德问题?
- 计算机只会返回Yes or No，但是我们知道实际授权过程中不只是这些，还需要审核判断，撤销等还涉及人类情感。AI化的机制，可能影响人类相互之间的信任关系，还有很多需要考虑。

What Can Political Philosophy Teach Us about Algorithmic Fairness?

1801210817 楚选耕

Introduction

Charging Plus-VIP a higher price.

Fair and Non-discriminatory

What does it mean for an **algorithmic decision-making system** to be “**fair**” or “**non-discriminatory**” in terms that can be operationalized?

Is it just a problem for moral and political philosophers?

Discrimination-aware Data Mining
Fair Machine Learning

Discrimination in Algorithm

If the possession of certain mental states by decision-makers is a necessary condition of a decision being discriminatory, one might argue that algorithmic decision-making systems can never be discriminatory as such, because such systems are incapable of possessing the relevant **mental states**.

But in fact there **IS** discrimination

Example:
Hiring Algorithm → Gender Race

Statistical Discrimination

Failing to Treat People as **Individuals**

Egalitarianism

Rights like voting in elections

The aim of egalitarianism might be **absolute equal distribution** of the good, rather than merely equality of opportunity to compete for it.

Rights to get a good job

When it comes to competition for social positions and economic goods, we may be concerned with ensuring **equality of opportunity** but less concerned about equality of outcome.

Luck Egalitarian

Inequalities are the result of circumstances outside an individual’s control

Being born with a debilitating health condition or being born into a culture in which one’s skin color results in systemically worse treatment.

Is these should be used as variables for analyzing?