Artificial Moral Advisors:

A New Perspective from Moral Psychology

Yuxin Liu, Adam Moore, Jamie Webb, and Shannon Vallor

The Fifth ACM/AAAI Conference on AI, Ethics, and Society (AIES '22) August 1-3, 2022, Oxford, United Kingdom







Artificial Moral Advisor (AMA)

"... a type of software that would give us moral advice more quickly and more efficiently than our brain could ever do, on the basis of moral criteria we input."

Giubilini & Savulescu (2018, p. 171)

A tool for moral AI enhancement

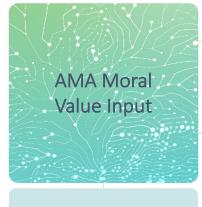
- Provides tailored moral advice based on preencoded moral values
- Relativist quasi-ideal observer

AMA Internal Configuration



Responding to AMA Moral Advice

AMA Internal Configuration



Updating or shifting of moral values

Lack of universal consensus

Conflicts between principles & the lack of exception-less universals

Dual-process of moral judgement



Varying cognitive representations underpinning judgement & decision-making

Lay theories of metalevel moral preferences

Responding to AMA Moral Advice

Passive acceptance?

Ignoring the AMA

Irreducibility & inescapability of moral decisions



Accepting/rejecting the AMA: responses to AMA are by nature human moral judgments

Motivated cognition & selective information processing

Moral degradation

Exploitation of a polarising AMA

Procedural elements of morality?

True Al Moral Enhancement?