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

Centre for **Technomoral Futures**

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

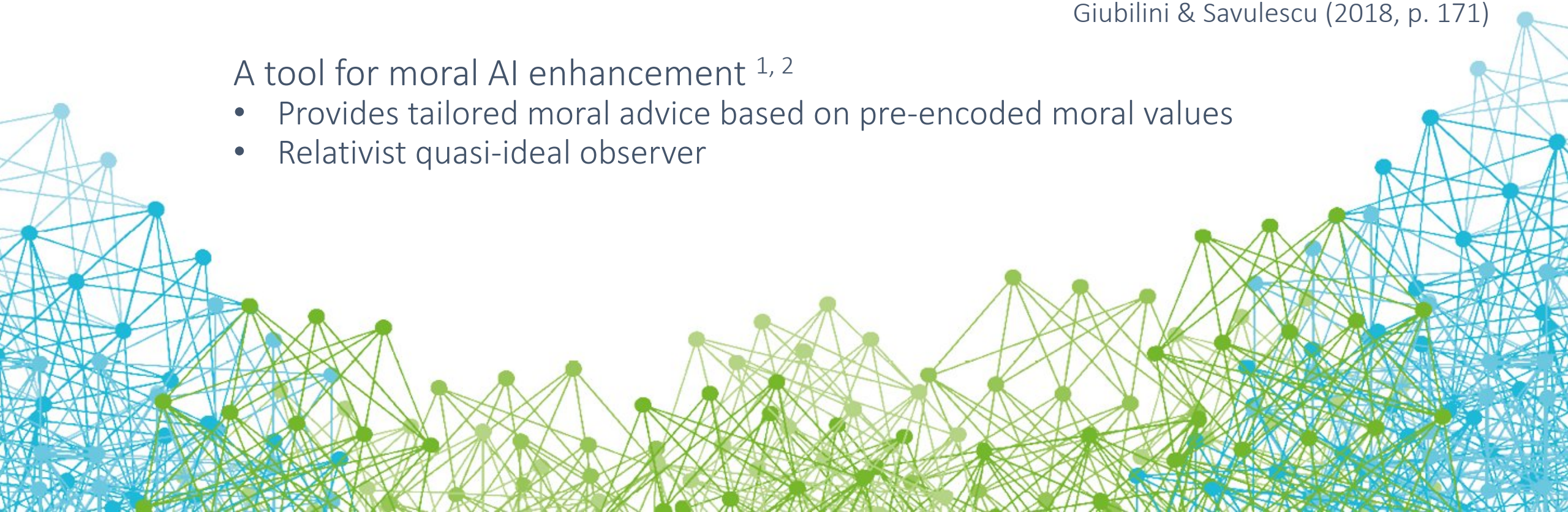
Artificial Moral Advisor (AMA)

"… 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."

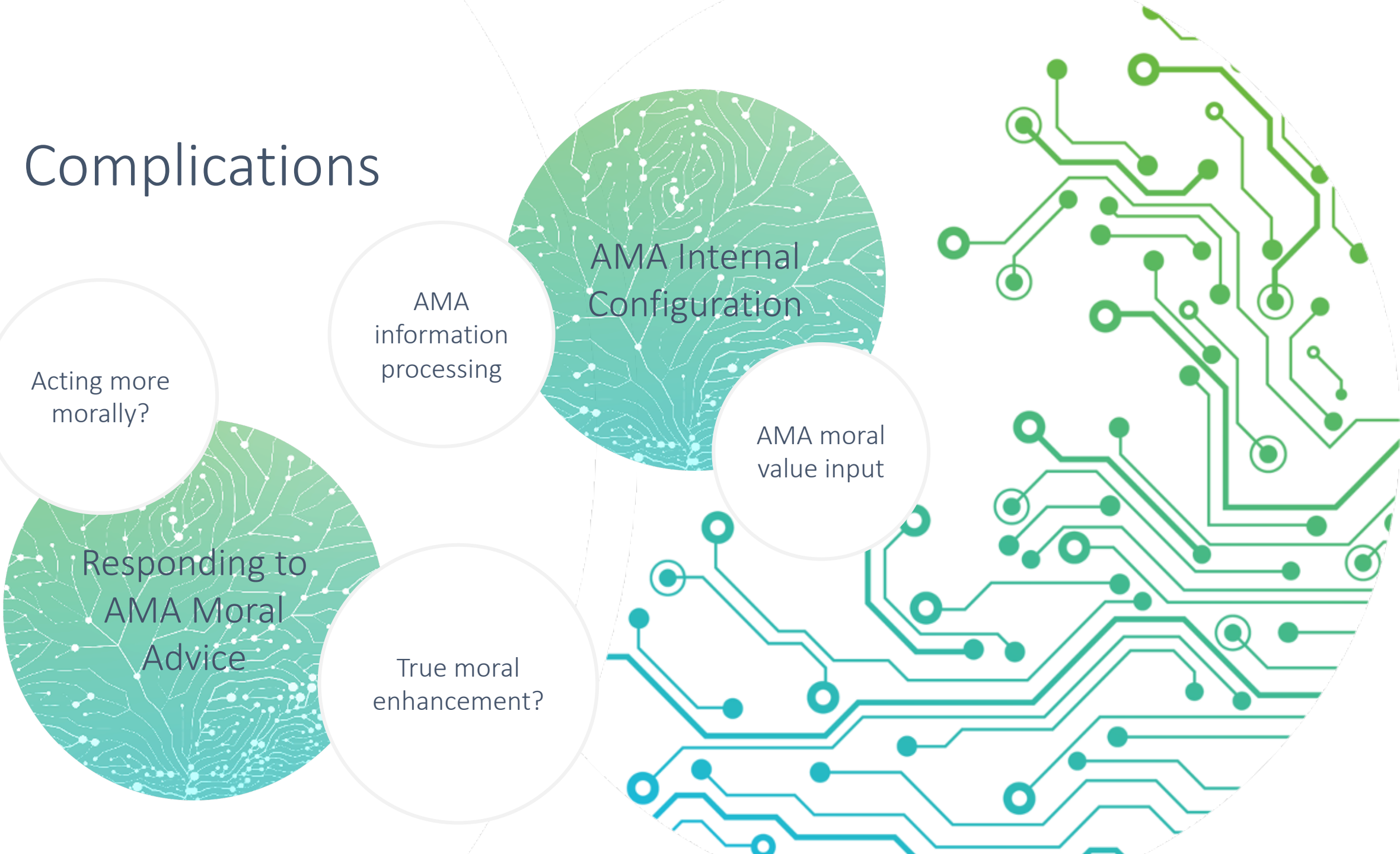
Giubilini & Savulescu (2018, p. 171)

A tool for moral AI enhancement [1, 2]

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

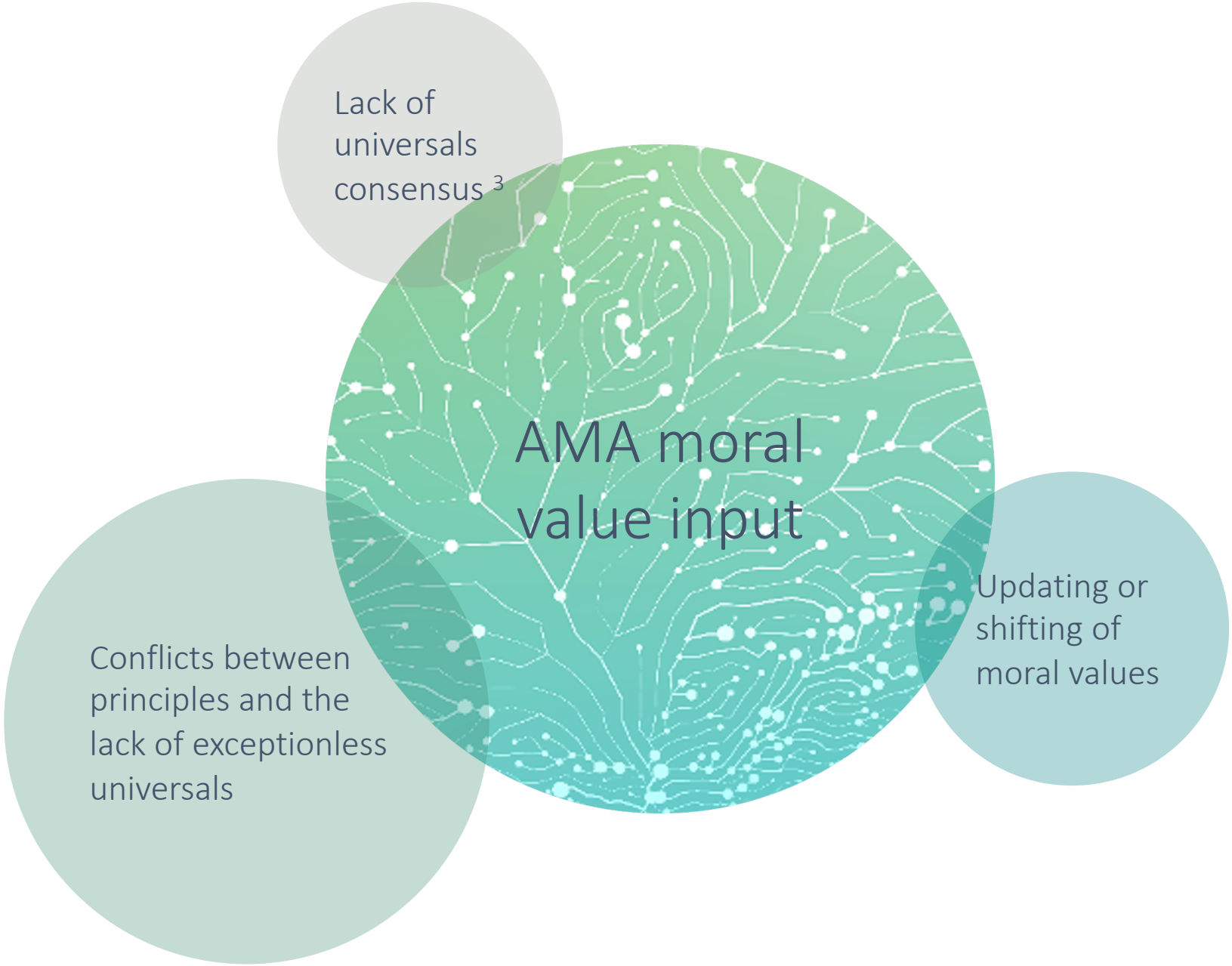


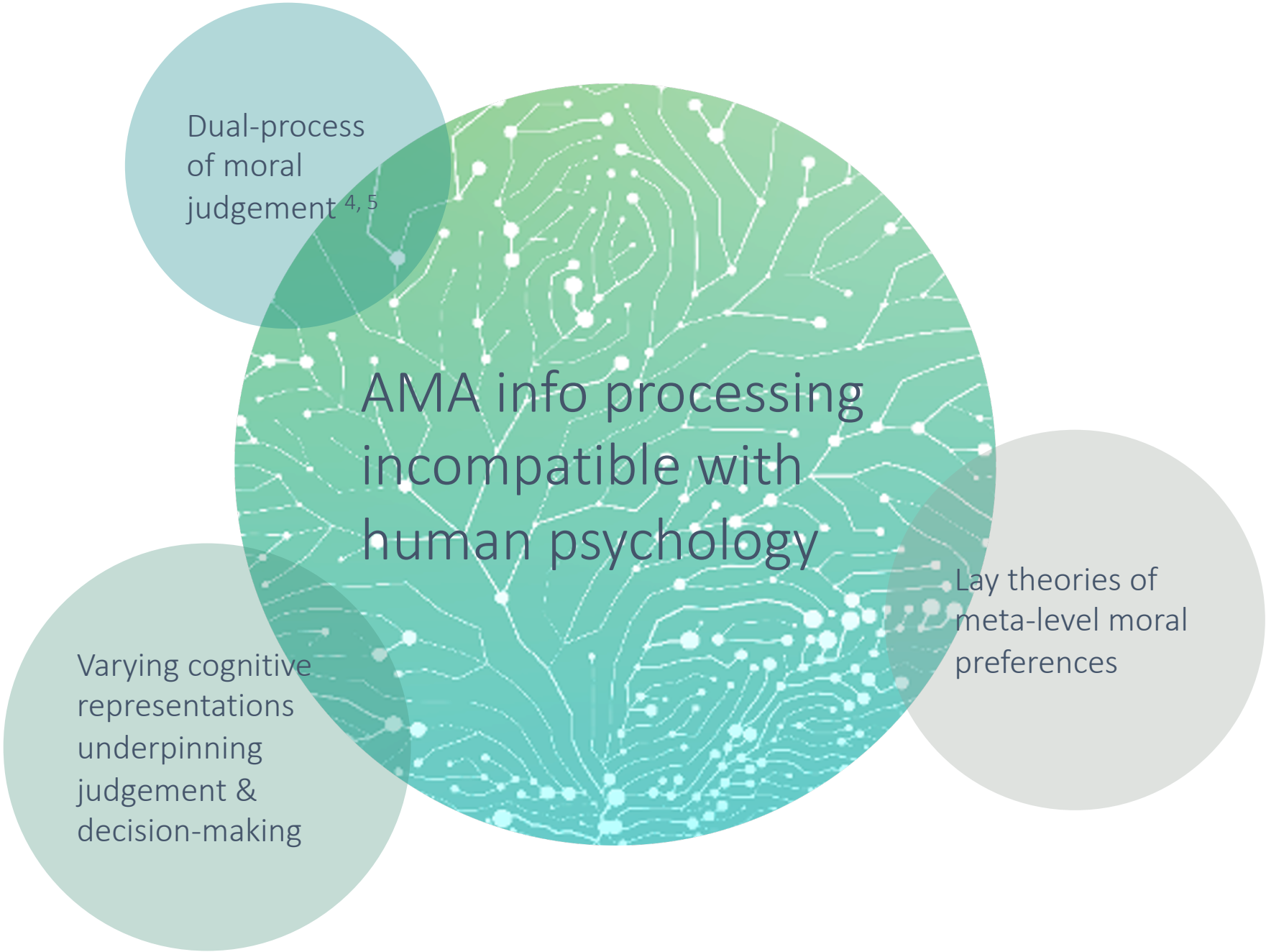
Complications



AMA Internal Configuration

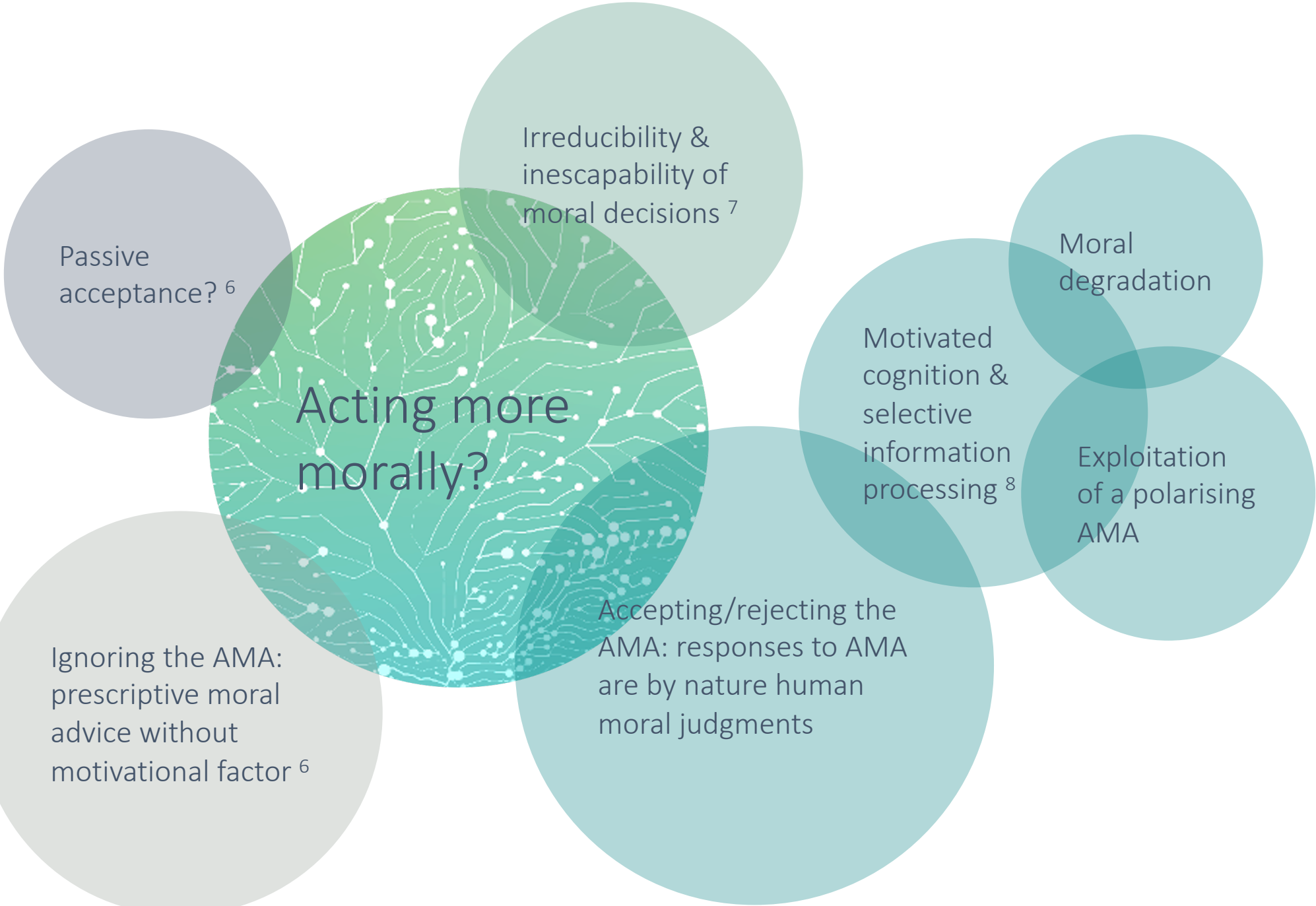
- AMA existing concerns: Moral value input
- AMA information processing: Incompatibility with human psychology

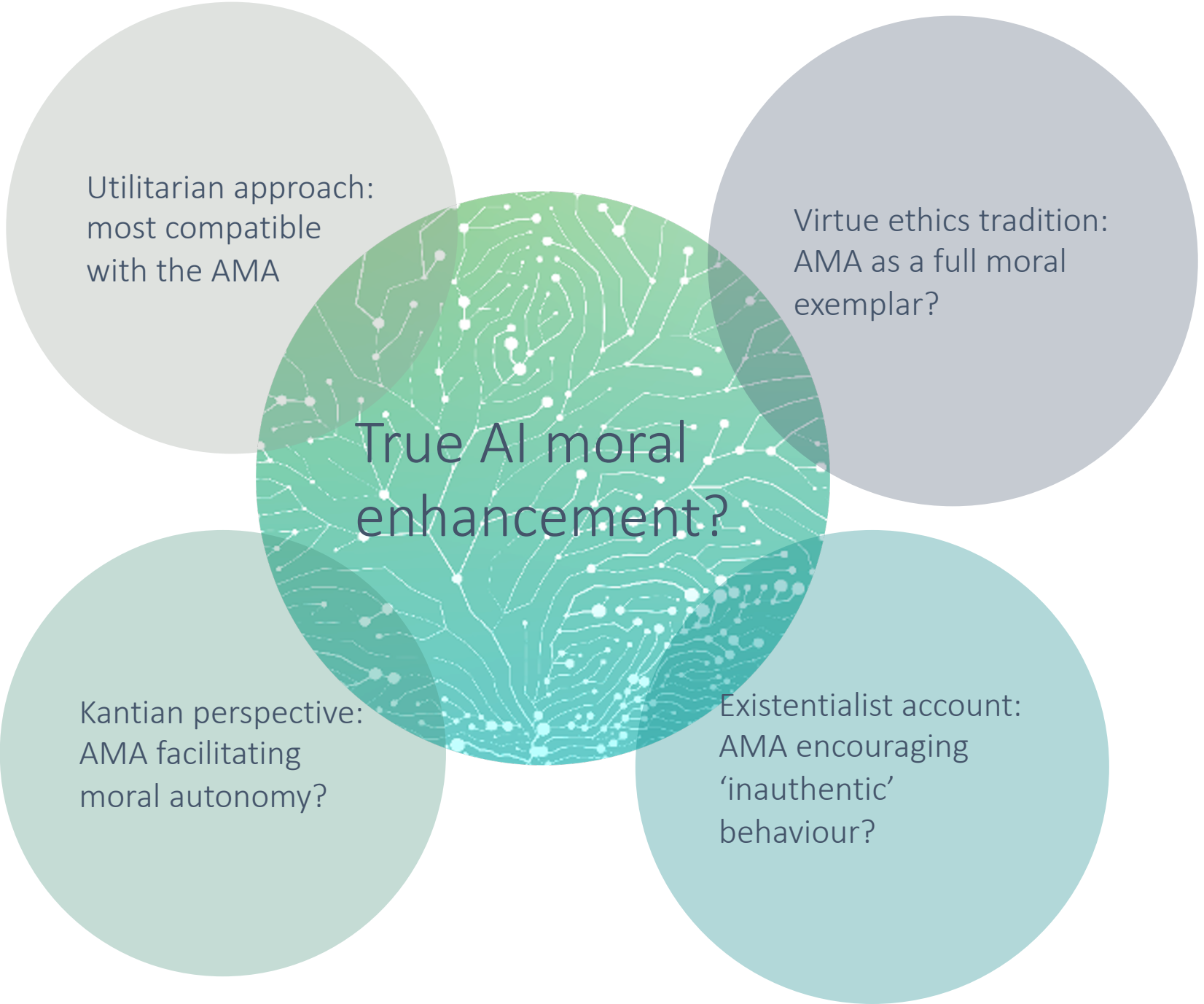

Lack of universals consensus [3]
AMA moral value input
Conflicts between principles and the lack of exceptionless universals
Updating or shifting of moral values


Dual-process of moral judgement [4, 5]
AMA info processing incompatible with human psychology
Varying cognitive representations underpinning judgement & decision-making
Lay theories of meta-level moral preferences

Responding to AMA Moral Advice

- Acting more morally?
- True AI moral enhancement?

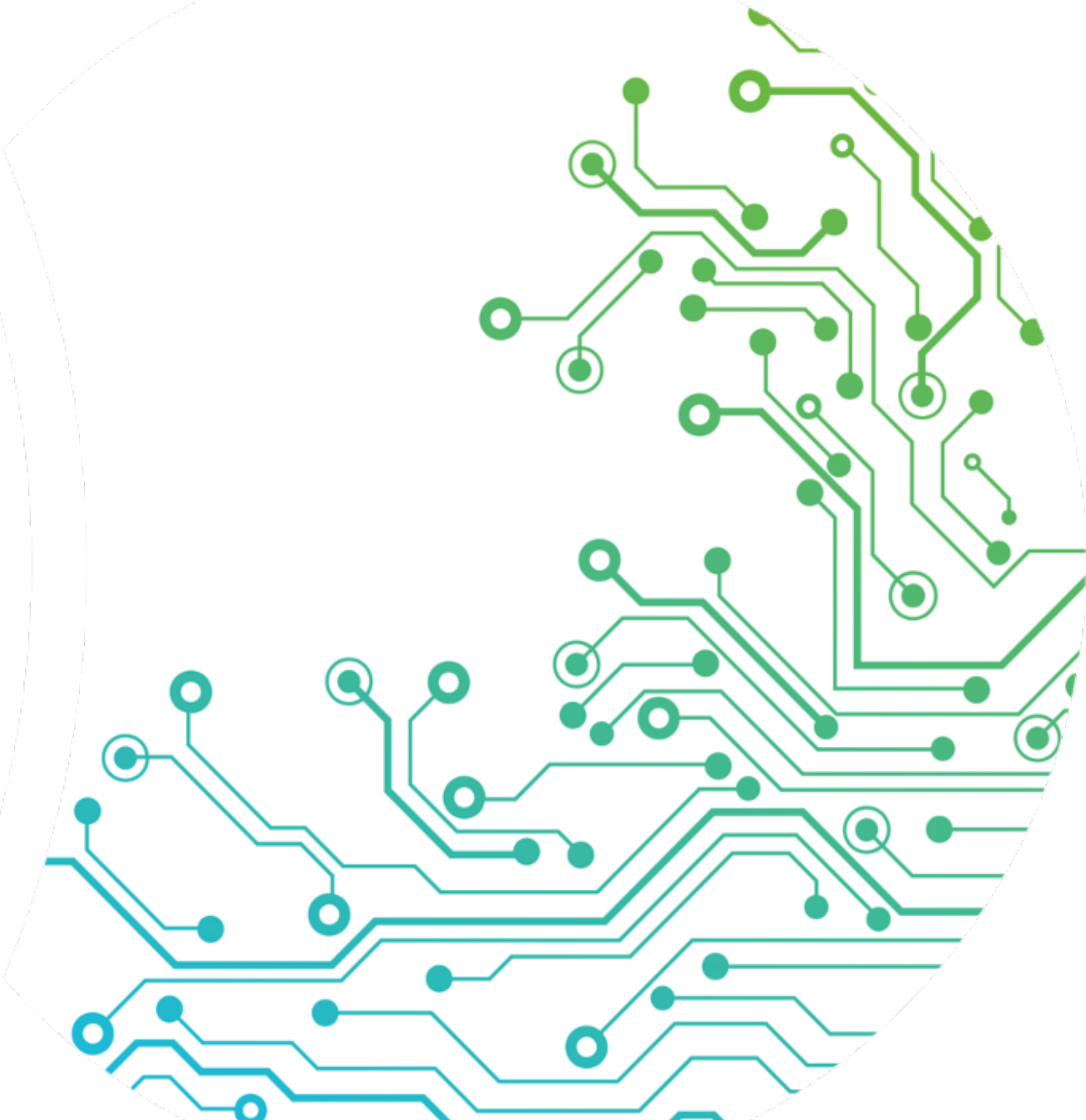

Passive acceptance? [6]
Irreducibility & inescapability of moral decisions [7]
Acting more morally?
Moral degradation
Motivated cognition & selective information processing [8]
Exploitation of a polarising AMA
Accepting/rejecting the AMA: responses to AMA are by nature human moral judgments
Ignoring the AMA: prescriptive moral advice without motivational factor [6]


Utilitarian approach: most compatible with the AMA
Virtue ethics tradition: AMA as a full moral exemplar?
True AI moral enhancement?
Kantian perspective: AMA facilitating moral autonomy?
Existentialist account: AMA encouraging ‘inauthentic’ behaviour?

Positive Use Case: AMA in Healthcare

Domain-specific AMA for clinicians/physicians?

Conclusions



References

1. Giubilini, A. and Savulescu, J. 2018. The artificial moral advisor. The “ideal observer” meets artificial intelligence. *Philosophy & Technology*. 31, 2 (Jun. 2018), 169–188. DOI:https://doi.org/10.1007/s13347-017-0285-z.
2. Savulescu, J. and Maslen, H. 2015. Moral enhancement and artificial intelligence: Moral AI? *Beyond artificial intelligence*. J. Romportl et al., eds. Springer International Publishing. 79–95.
3. van Wynsberghe, A. and Robbins, S. 2019. Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*. 25, 3 (Jun. 2019), 719–735. DOI:https://doi.org/10.1007/s11948-018-0030-8.
4. Crockett, M.J. 2013. Models of morality. *Trends in Cognitive Sciences*. 17, 8 (Aug. 2013), 363–366. DOI:https://doi.org/10.1016/j.tics.2013.06.005.
5. Crockett, M.J. et al. 2021. The relational logic of moral inference. *Advances in experimental social psychology*. Elsevier. 1–64.
6. Lara, F. and Deckers, J. 2020. Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics*. 13, 3 (Oct. 2020), 275–287. DOI:https://doi.org/10.1007/s12152-019-09401-y.
7. Sartre, J.-P. 1946. *Existentialism is a humanism*.
8. Liu, Y. and Moore, A. 2022. A Bayesian multilevel analysis of belief alignment effect predicting human moral intuitions of artificial intelligence judgements [Forthcoming]. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (Toronto, Canada, 2022).

Centre for **Technomoral Futures**

https://www.technomoralfutures.uk/

https://www.technomoralfutures.uk/phd-students/yuxin-liu

https://www.technomoralfutures.uk/phd-students/jamie-webb

ctf@ed.ac.uk

yliu3310@ed.ac.uk

jamie.webb@ed.ac.uk

@CentreTMFutures

@_yuxinl_

@JamieDWebb