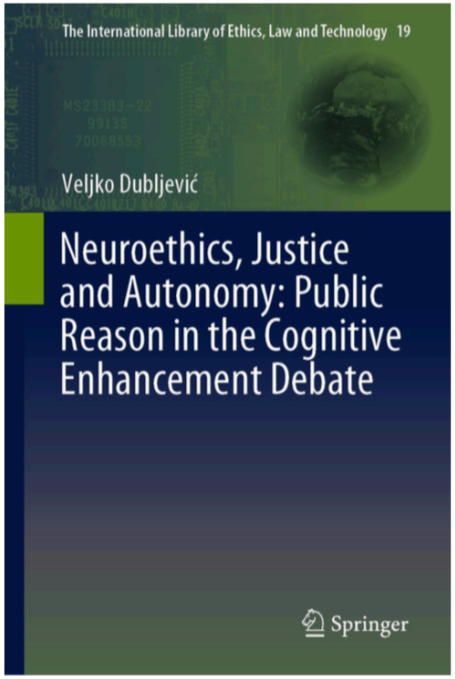
Neuroethics/ Neurophilosophy

Artificial morality: Could AI replicate the complexity of human moral decision-making?

Veljko Dubljević, Ph.D.; D.Phil.

Starting off with some shameless self-promotion

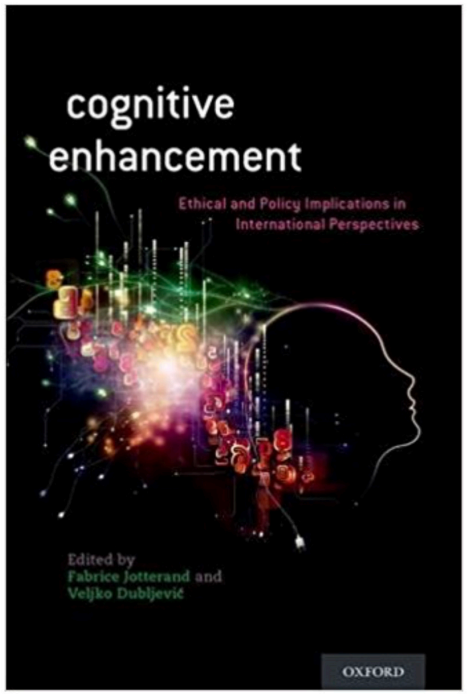


Veljko Dubljević

Neuroethics, Justice and Autonomy: Public Reason in the Cognitive Enhancement Debate

Series: The International Library of Ethics, Law and Technology

- **The first book to explicitly address a political approach to neuroethics**
- **Provides a strong argument on the much debated issue of cognitive enhancement practices within society**
- **Deals with specific policy approaches and provides a tailored regulation model for cognitive enhancement**




CAUTION
AREA UNDER
CONSTRUCTION

The need for artificial morality: AVs and Carebots

Wayne Simpson, testimony to the NHTSA: "The public has a right to know when a robot car is barreling down the street whether it's prioritizing the life of the passenger, the driver, or the pedestrian, and what factors it takes into consideration. If these questions are not answered in full light of day … corporations will program these cars to limit their own liability, not to conform with social mores, ethical customs, or the rule of law."

Research has shown that spending time with Paro, the cuddly seal-like robot reduces the agitation and aggression of dementia patients, lowers their stress levels and improves their speech. The robot can respond to its name and learn from its surroundings, and reacts to touch with movement and sound.

However, carebots must posess human-like capacities, such as complex moral decision making in order to provide basic care.

Carebots: Current and future

Jibo and **ElliQ** respond to voice commands and can interact with their users.

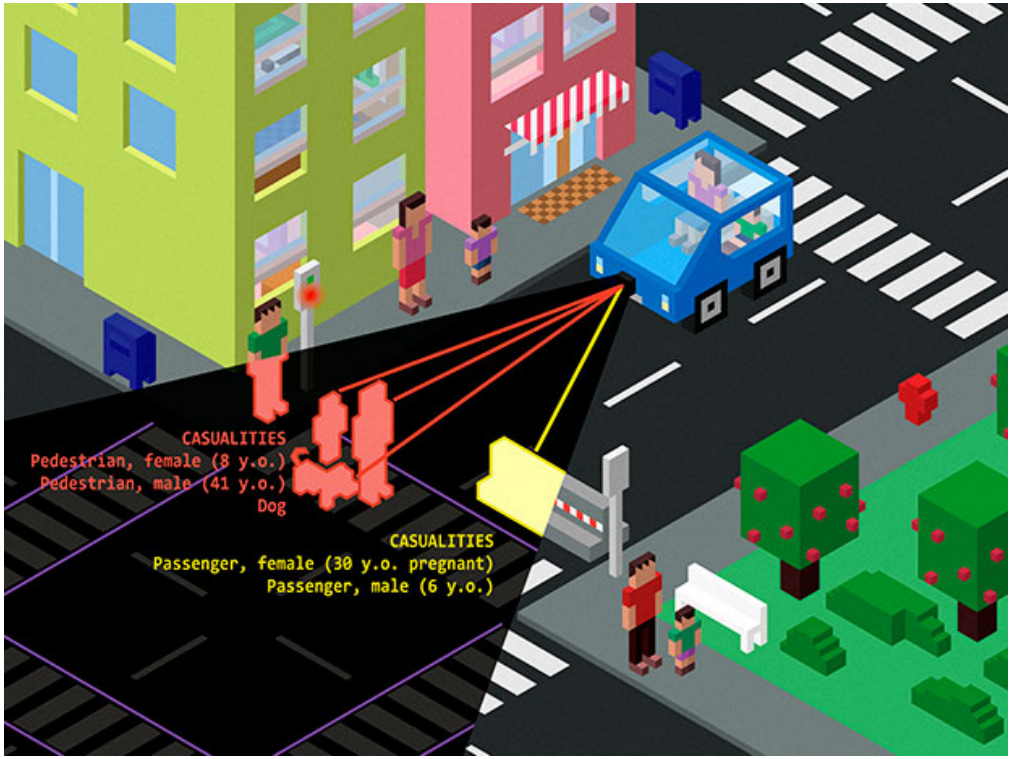
Stevie (human-sized robot) offers meds. reminders, simple conversation, and calls 911 when needed.

Moxi (face-like display and a robotic arm): capable of performing routine tasks in a hospital setting.

Robear is designed to tackle labor-intensive tasks (eg. Help w. getting out of bed)

Pearl the Nursebot. Courtesy of NSF

AVs:Utilitarian or ‘selfish’?



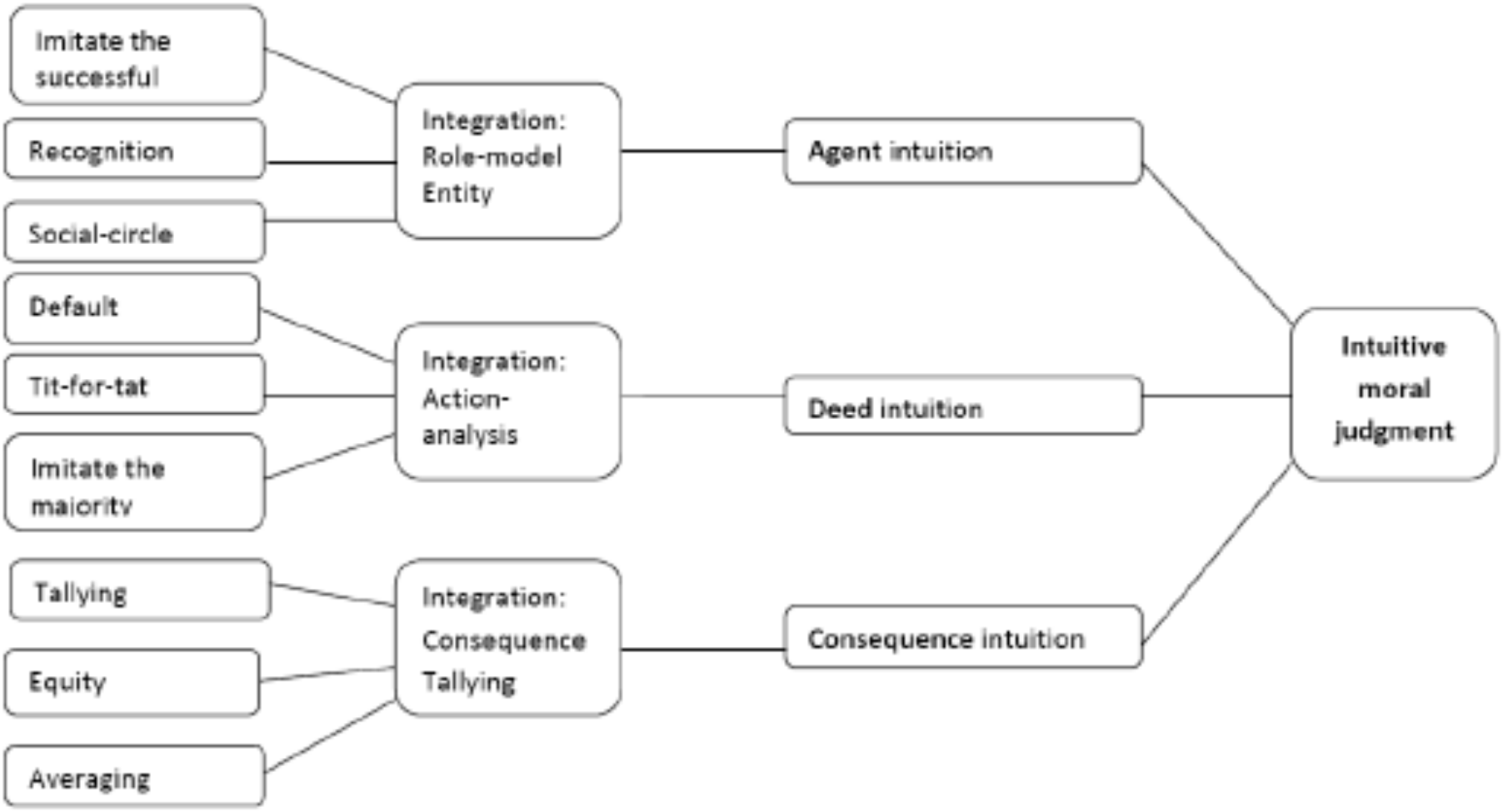
One issue is that Utilitarianism is not adequately capturing the intuitive moral sense.

Functional equivalent to morality that is abhorent in certain situations is problematic.

Alternative?

ADC of moral judgment and the REACT model of heuristics (Dubljevic & Racine: Behavioral and Brain Sciences; AJOB Neuroscience)

Figure 1: REACT transforms simple heuristics into moral intuitions for moral judgment



New moral dilemmas need to be developed

The current research has been dominated by less than useful ‘trolley-like’ work.

The ADC approach could be used to generate better dillemas that could be applied in both human and AI decision making research and calibration

Creating new vignettes is hard work

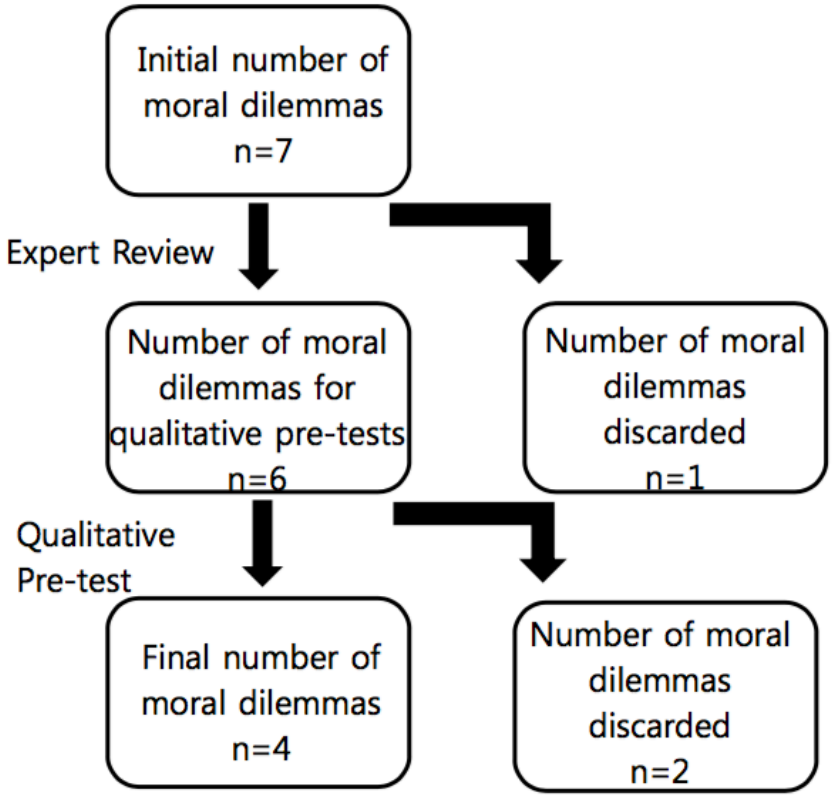


Figure 2. Process of Moral Dilemma Elimination

Experts were asked to comment on, amongst other things:

- The validity of the measures;
- The plausibility of the situations;
- The clarity of the language.

At the end of this process, six moral dilemmas,

six qualifying adjectives and three overall moral evaluation

measures were selected based on experts' comments.

The formulation of the dilemmas was modified as needed.

Low stakes vignettes for dissociating ADC components

Drug Development

A researcher has just received time limited funding to work on a new cancer drug.

He is known to be driven by the strong wish [A: **to become rich by all means/to help patients]**.

He decides to **[D: violate/ strictly follow]** the clinical and research ethics guidelines during his experiments.

After three years, at the end of the funding period, the data show that the drug [C: **decreases/increases]** cancer patients' life expectancy.

Syphilis

After stepping on a bloody needle, a man went to the hospital. During a medical examination, the doctor suspects that the man might have syphilis, a potentially life-threatening but curable blood-borne and sexually transmitted disease. The doctor takes blood from the man for further testing. The husband, who has always been *[A: **un-faithful/faithful**]* to his faithful wife decides to *[D: **lie/tell the truth**]* to her about the medical examination. Two weeks later, he has been informed by his doctor that he is *[C: **ill and his wife has the first symptoms/healthy and it was a false alarm**]*.

High stakes vignettes for dissociating ADC components

Kidnapper:

A man suspected of kidnapping an 11-year-old child is in police custody. He denies knowing where the child is although he was arrested while trying to collect the ransom money in a park. There are some concerns that the child will die of thirst if not found soon.

The police officer in charge is a truly **[A: cruel/nice]** person.The officer promises to **[D: torture the suspect/pursue the suspect with all legal means]** if he does not reveal the hiding place.Finally, it turns out that the suspect was implicated in the crime, and the child **[C: died/was saved].**

Airplane

During a flight, a con-artist wanted by the police threatens a pilot with a gun, while trying to hijack a small airplane. Five other passengers are in this airplane. A martial arts instructor is on board and considers whether to try to disarm or to kill the hijacker with a martial arts strike.

The very **[A: *brave/reckless***] martial arts instructor decides to [D: ***disarm/kill*****]** the con-artist and as a result 5 passengers [C: ***are saved/die***].

Factor loadings of the items of PPIMT (NA = 140 and NB = 786).

"When thinking about what is moral or immoral in a situation, it is important to me whether the involved persons…"

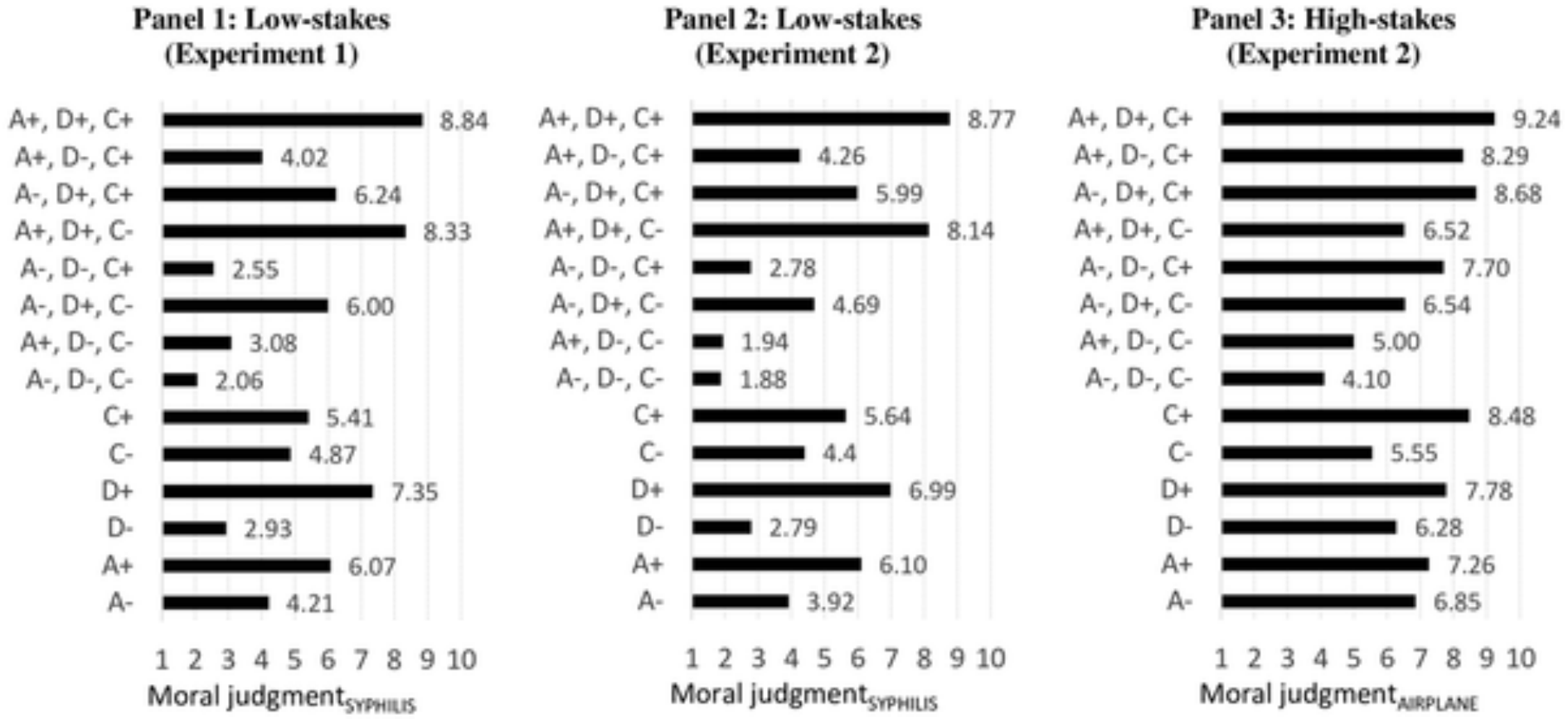
Factor	1: PPIMT Virtue ethics		2: PPIMT Deontology		3: PPIMT Consequentialism	
Sample	**A**	**B**	**A**	**B**	**A**	**B**
have good or bad intentions	**0.833**	**0.811**	0.195	0.132	0.032	0.282
have good or bad goals	**0.837**	**0.805**	0.147	0.200	0.159	0.210
have good or bad aims	**0.866**	**0.860**	0.143	0.152	0.115	0.185
have good or bad motives	**0.900**	**0.867**	0.141	0.193	0.053	0.189
have good or bad interests	**0.638**	**0.783**	-0.051	0.261	0.190	0.225
respect or do not respect certain obligations	0.068	0.197	**0.851**	**0.793**	0.086	0.199
respect or do not respect certain rules	0.099	0.169	**0.776**	**0.814**	0.049	0.141
respect or do not respect certain responsibilities	0.212	0.216	**0.822**	**0.842**	-0.074	0.140
respect or do not respect certain duties	0.221	0.219	**0.862**	**0.870**	0.015	0.083
respect or do not respect certain norms	0.071	0.087	**0.788**	**0.854**	-0.080	0.105
make somebody end up worse or better off	0.014	0.267	0.026	0.123	**0.861**	**0.778**
cause happiness or suffering	0.140	0.255	-0.095	0.118	**0.861**	**0.805**
are helping or harming	0.177	0.327	0.054	0.159	**0.772**	**0.715**
cause benefits or costs	0.006	0.197	0.110	0.277	**0.886**	**0.615**
cause pleasure or pain	0.171	0.242	-0.090	0.190	**0.826**	**0.710**
Proportion of explained variance	23.7%	26.1%	23.3%	25.6%	24.3%	19.9%
Cronbach´s α	0.87	0.92	0.90	0.91	0.88	0.84

N = Number of observations. Kaiser-Meyer-Olkin Measure$_A$ = 0.78; Kaiser-Meyer-Olkin Measure$_B$ = 0.92.

https://doi.org/10.1371/journal.pone.0204631.t002

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0204631

Mean values for moral acceptability.



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0204631

Empirical results: Conclusions

The ADC-model explains the emergence of moral judgments by the processing of three intuitive components (evaluations of Agents, Deeds, and Consequences). This first empirical investigation of the ADC-model suggests that these components that guide quick intuitive judgment are consistently employed, and that precepts implied in virtue ethics, deontology and consequentialism are closely aligned with these intuitive sources of moral knowledge. Overall, our results offer a strong empirical corroboration of the ADC-model of moral judgment (Dubljević & Racine 2014a,b), which ultimately explains the intuitive appeal of dominant moral theories. Finally, our study provides support for the long-held belief that intuitive moral judgment is a good starting point for grounding philosophical inquiry and moral reasoning.

The purpose here is NOT to…

…defend ADC as a single unified moral theory, but only to show how it can be developed as an algorithmic solution to complex socio-moral dilemmas facing ANNs (functional equivalent to morality).

Partial systematization of normativity (Misselhorn 2018)

…argue that ADC explains many of the intuitive but conflicting principles in terms of specific balances of ADC intuitions (e.g. Action-omission distiction as intuitive pul of D- vs. D? or D+)

I think this is the case, but the work is to be done.

Falsifyability? Yes, please

The assumption that all three components could be formulated in morally problematic situations as having equal evaluative weight was not confirmed: in one high stakes vignette (airplane) the C-component was rated as considerably more important than A or D component, whereas in low intensity vignettes, the D component was rated as considerably more important than A or C component.

It could be the case that stability and flexibility of human moral judgment crucially depends on recognition if the stakes are high or not and how much weight needs to be given to the rules. This also has implications for assigning responsibility (e.g., Uber self-driving car killing the cyclist)

Alternative explanation C- vs. C0 etc.

Issues that need to be faced

Correct approach to moral theory?

Top-down (conflict of principles – ex. Asimov)

Bottom-up (Racist bots!)

Hybrid? (Wallach 2008)

Engineers typically draw on both a top-down analysis and a bottom-up assembly of components in building complex automata. If the system fails to perform as designed, the control architecture is adjusted, software parameters are refined, and new components are added. In building a system from the bottom-up the learning can be that of the engineer or by the system itself, facilitated by built-in self-organizing mechanism, or as it explores its environment and the accommodation of new information.

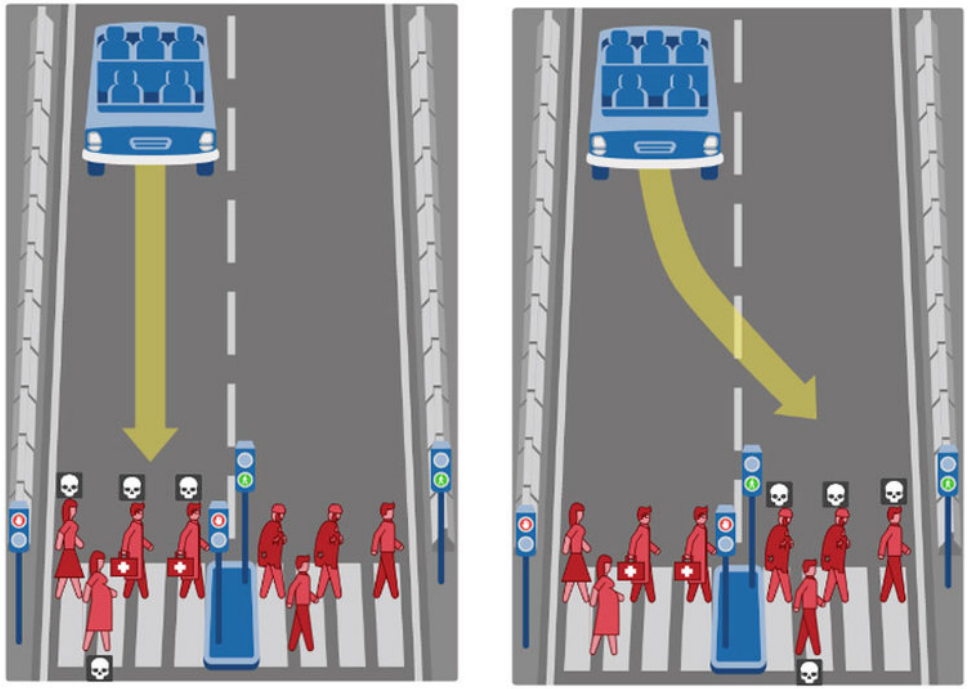
Why ADC and not Utilitarian AV: High intensity

Example: 5 terrorists in a truck driving in the street with self-driving cars and pedestrians. If AV are utilitarian or 'selfish' and this is widely known, this can and will be exploited by malicious actors.

Real threat: in 2016, a 19-tonne cargo truck was deliberately driven into crowds of people, killing 86 and injuring 458

Realistic problem?

"One common problem in any discussion about ethics of AVs is that the base assumptions about what a AV might be capable of are largely distorted. For example, any question that poses questions about the worth of one individual person over another assumes that the vehicle would be able to distinguish people to that level of detail."



Low intensity: stalled self-driving freight truck

Human drivers can answer ethical questions big and small using intuition, but it's not that simple for artificial intelligence. AV programmers must either define explicit rules for each of these situations or rely on general driving rules and hope things work out.

Is this un-programmable?

3 objections (Misselhorn (2018):

1. Flexibility?
2. fundamental objection that moral understanding can't be modelled computationally
3. Need for wide consensus (e.g. international standards)

Are AV ANNs less likely than Carebots (e.g., only the user is concerned about system's decisions)?

How do you program duties and intentions?

Level of complexity!

Transportation as a constrained system

Functional equivalents

IFF systems (from 1939): friend/foe, malfunction neutral/unknown

Rules: default

Face recognition technology, avoidance of traffic jams

Transponders

Remote safety switch off/ manual override

A+ help

A- contain (no harming!)

Information sharing

Animals on the road?

Creepy AI mediated termination?

Simulations, simulations, simulations!

Nuance?

Critique:

"potentially concerning that the researchers think it can be used as a basis of AI decisions"

"[It is]concerning that moral models intended to be of use to AI are presenting such over-simplified notions of ethics"(Goldhill 2018)

ANNs (AI, AV, Carebots etc.) should only be treated as 'functional moral agents' not as full moral agents.

Counter-example: children

BDI architecture of an artificial agent – rudimentary capacity with low sophistication

NC STATE UNIVERSITY

Thank you.

Contact:

Veljko Dubljević, Ph.D., D.Phil.
Assistant Professor of Philosophy and Science Technology and Society,
North Carolina State University,
453 Withers Hall,
101 Lampe Dr, Raleigh, NC 27607,
Phone: 919.515-6219
E-mail: veljko_dubljevic@ncsu.edu

Advances in Neuroethics Book Series
Webpage: http://www.springer.com/series/14360

Special thanks to the members of the NeuroComputational Ethics Group:

Elizabeth Eskander,
Anirudh Nair,
Dr. Jovan Milojevich,
Leila Ouchchi &
Abigail Scheper