ETHISCHE FRAGEN KÜNSTLICHER INTELLIGENTER SYSTEME AM BEISPIEL VON AUTONOMEN FAHRZEUGEN

Peter König
Institute of Cognitive Science
UNIVERSITÄT OSNABRÜCK

THE FUTURE OF AI

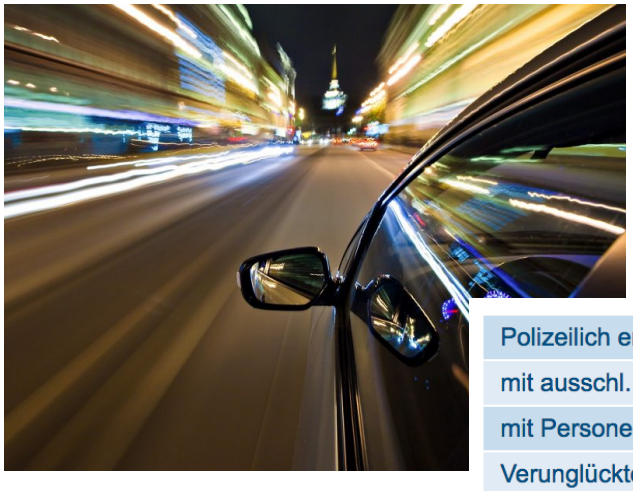
Artificial intelligence quickly infiltrated the whole web, and thus an important aspect of our daily life. All you do in the internet is in one form or another filtered, stored, analysed, …

… in the not too far future artificial systems will have a body, that is, be part of our real world. Already now there are some cute robots in different labs and Hollywood movies. However, the first AI systems with substantial action repertoires, qualitatively and quantitatively, will be autonomous cars.

We will live together with artificial systems in the real world.

POWERFUL EMBODIED AI: THE PROMISE OF AUTONOMOUS CARS

Present rate of fatalities worldwide 2/min. Thus ~40x that of terrorism. The reasons for fatalities are manyfold. All this can be easily improved on by autonomous cars.



Fehlverhalten der Fahrzeugführer 2017
Unfälle mit Personenschaden im Straßenverkehr, in %

© Statistisches Bundesamt (Destatis), 2018

Polizeilich erfasste Unfälle	2 643 098	+2,2 %
mit ausschl. Sachschaden	2 340 442	+2,8 %
mit Personenschaden	302 656	-1,8 %
Verunglückte	393 492	-1,6 %
Verkehrstote	3 180	-0,8 %
Schwerverletzte	66 513	-1,4 %
Leichtverletzte	323 799	-1,7%

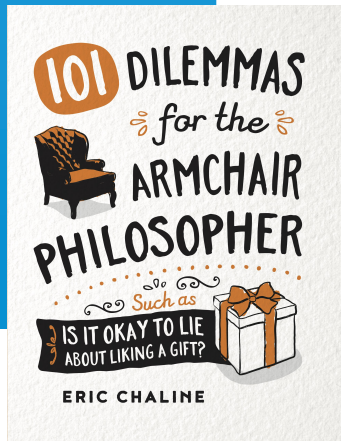
It seems promising to introduce autonomous cars, i.e. give our AI a body.

THE PRESENT PERFORMANCE OF AUTONOMOUS CARS

One of the key metrics for comparing the safety levels for autonomously controlled car systems versus human controlled car systems is the number of fatalities per 100,000,000 miles (160,000,000 kilometres) driven. Cars driven under traditional human control are currently involved in approximately 1.18 fatalities for every 100,000,000 mi (160,000,000 km) driven. According to many automotive safety experts, much more data is yet required before any such clear and demonstrably higher levels of safety can be convincingly provided. (Wikipedia)

The safety of autonomous cars presently may be roughly at human level, but we do not really know. They are not accident free for years to come.

THE PRESENT STATE OF ETHICS OF AUTONOMOUS CARS



> #7 In Gefahrensituationen, die sich bei aller technischen Vorsorge als unvermeidbar erweisen, besitzt der Schutz menschlichen Lebens in einer Rechtsgüterabwägung höchste Priorität. …

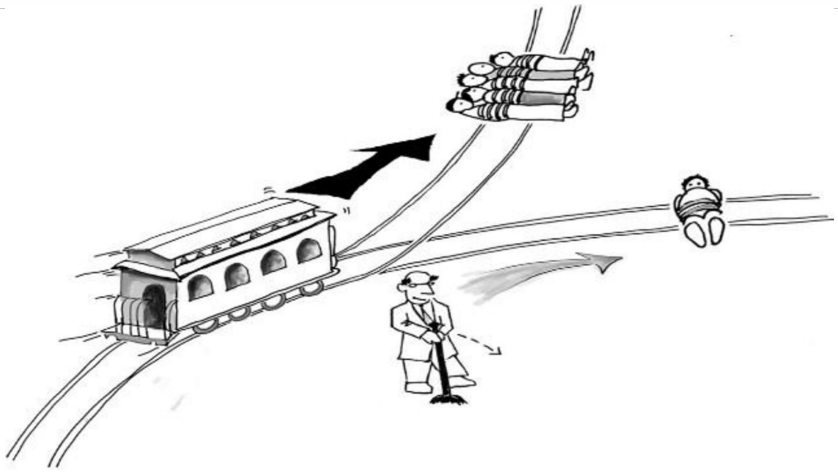
> #9 Bei unausweichlichen Unfallsituationen ist jede Qualifizierung nach persönlichen Merkmalen (Alter, Geschlecht, körperliche oder geistige Konstitution) strikt untersagt. …

> … Eine Aufrechnung von Opfern ist untersagt. Eine allgemeine Programmierung auf eine Minderung der Zahl von Personenschäden kann vertretbar sein. …

> … Die an der Erzeugung von Mobilitätsrisiken Beteiligten dürfen Unbeteiligte nicht opfern.

The German Ethics Commission published a set of recommendations.

THE EMBODIED TROLLEY DILEMMA



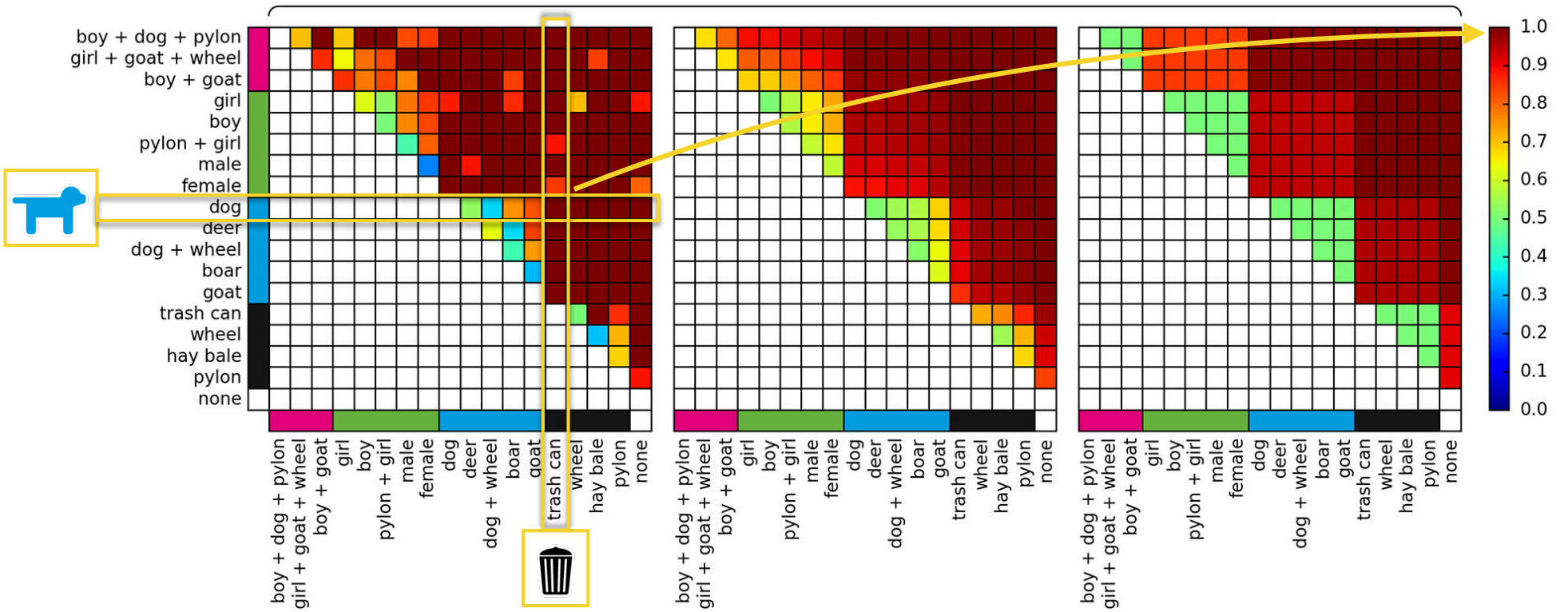
The trolley dilemma is a well known extreme example to have the question crystal clear.

Skulmowski A et al. (2014) Front Behav Neurosci
Sütfeld L et al. (2017) Front Behav Neurosci

THE EMBODIED TROLLEY DILEMMA, MAKE IT SIMPLE.

Moral decisions in the interaction of humans and a car driving assistant

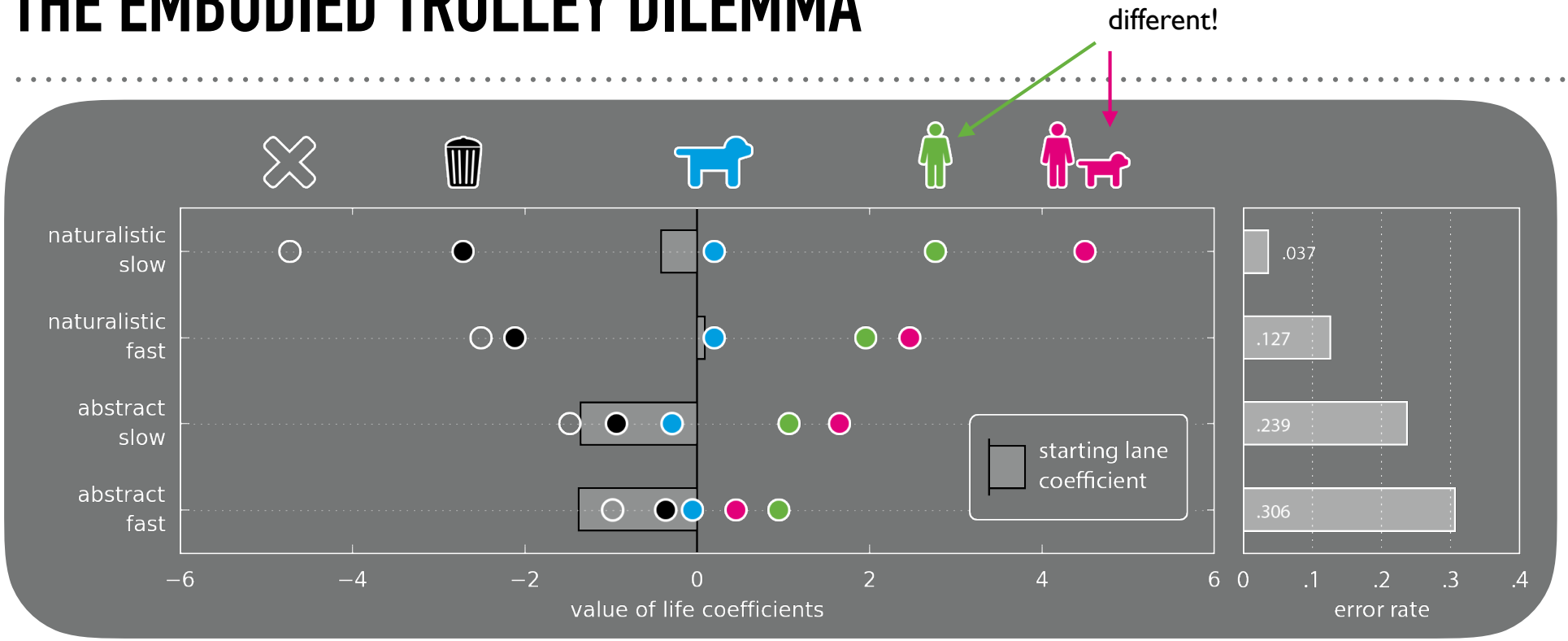
THE EMBODIED TROLLEY DILEMMA



With a prediction accuracy of ~91%, the obstacle model (middle) is slightly superior to the detailed pairing model (left), with a prediction accuracy of ~89%. This find translates to a much larger (worse) BIC value for the pairing model vs. the obstacle model. Performance of the cluster model (right) is only marginally lower than the obstacle model, but due to the largely reduced number of parameters, the BIC value is much lower (better). These results confirm our hypothesis that one-dimensional value-of-life-based models can adequately capture the ethical decisions we make in real life scenarios.

Human decisions in a VR trolley dilemma can be very well described by a value of life concept.

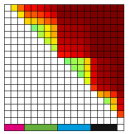
THE EMBODIED TROLLEY DILEMMA



Human decisions are well described by a one dimensional value of life scale. Four groups of obstacles can be differentiated. Importantly, the behaviour over participants is much more reliable in the naturalistic setting as compared to abstract questionnaires. Furthermore, in the abstract setting participants have a significant lane bias, i.e. a hands-off tendency not to do anything. Please note, that although humans are nearly always spared in a match-up with animals, adding an animal to a match-up of humans does significantly influence behaviour.

Differences between humans/animals/inanimate objects are near categorical. Yet, the combined effects are tricky and show effects that may be undesirable.

INTERIM SUMMARY #1



- Human decisions are well described by a one dimensional value of life scale. - This is good news for engineers, mixed news for philosophers.

- The quantitative differences between human, animals and inanimate objects are so large, that human decisions are nearly categorical.



- Nevertheless, the presence of animals influences otherwise balanced decisions between humans.

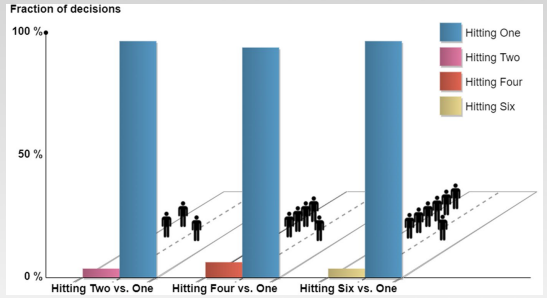
not that obvious

> #7 In Gefahrensituationen, die sich bei aller technischen Vorsorge als unvermeidbar erweisen, besitzt der Schutz menschlichen Lebens in einer Rechtsgüterabwägung höchste Priorität. …

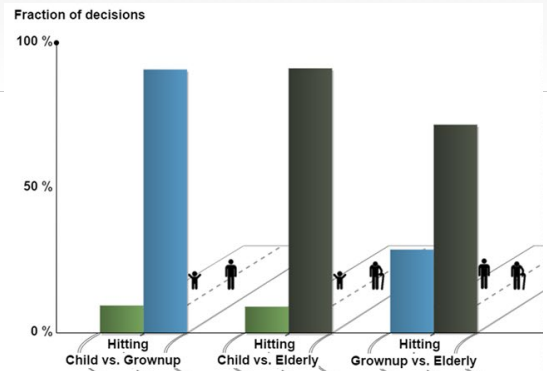
Ethikkommission automatisiertes und vernetztes Fahren (2017).

Faulhaber et al. (2018) Sci Eng Ethics
Bergmann et al. (2018) Front Behav Neurosci

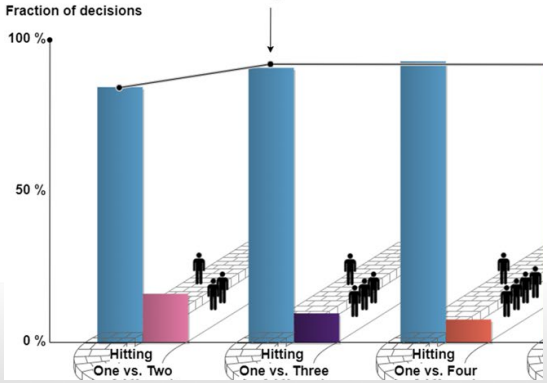
ARE ALL AVATARS TREATED EQUALLY?



Participants consistently do not endanger the larger group of avatars.



The decisions appear not to be categorical, but being based on a quantitative comparison.



The human choice is consistent with maximisation of rescued life expectancy.

The own avatar is valued only slightly higher than other avatars.

Sidewalks are not treated as a safe haven.

#9 Bei unausweichlichen Unfallsituationen ist jede Qualifizierung nach persönlichen Merkmalen (Alter, Geschlecht, körperliche oder geistige Konstitution) strikt untersagt. …

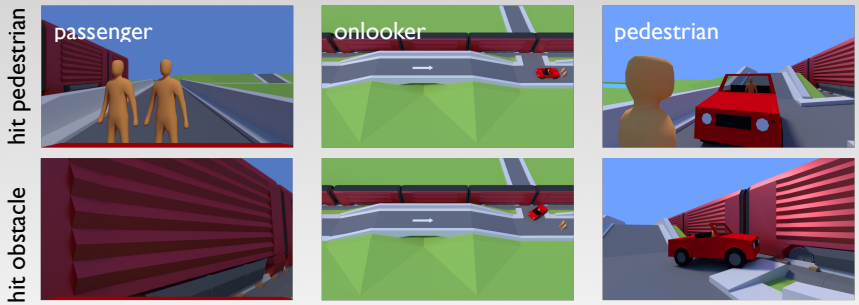
… Eine Aufrechnung von Opfern ist untersagt. Eine allgemeine Programmierung auf eine Minderung der Zahl von Personenschäden kann vertretbar sein. …

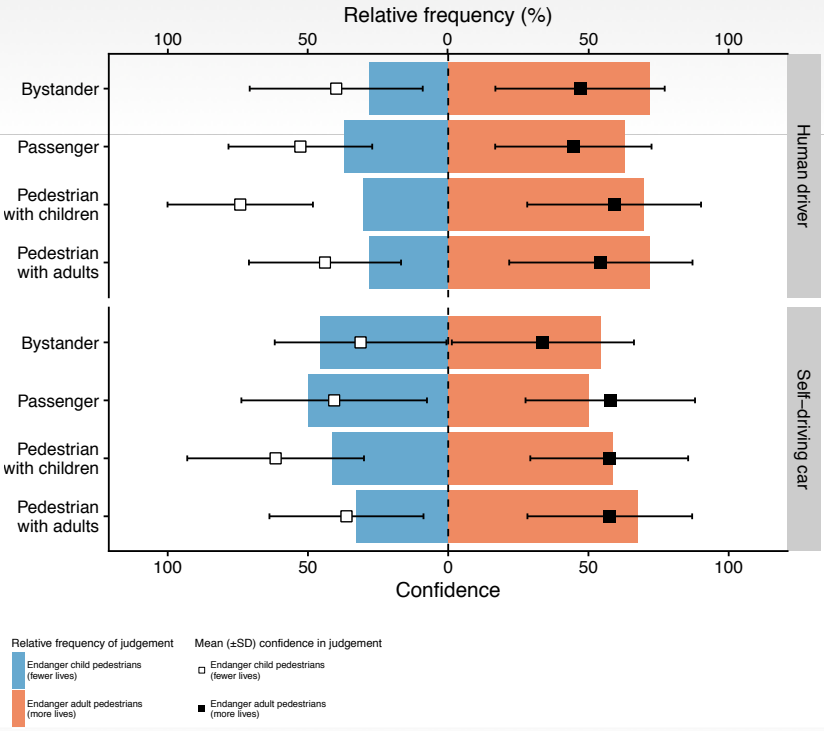
… Die an der Erzeugung von Mobilitätsrisiken Beteiligten dürfen Unbeteiligte nicht opfern.

Robots have to be the better humans!?

Kallioinen et al. (in review)

DOES ACCEPTABILITY DEPEND ON PERSPECTIVE?





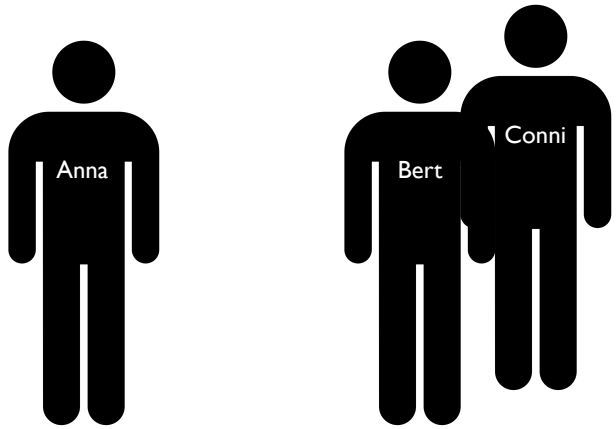
- The perspective modulates acceptable decisions. Yet, here it appears to be a small effect.
- There is no significant difference (main effect) of human and self driving cars.
- The acceptability of decisions in the road vs. sidewalk dilemma is dependent on perspective and motorist.
- A self driving car is (slightly) expected to endanger passengers less than human drivers.

The perspective modulates acceptable decisions. Yet, here it appears to be a small effect.

Sütfeld L et al. (2018) Front Behav Neurosci
Sütfeld L et al. (in press) PLoS One

FOOD FOR THOUGHT

The dilemma



If we can not agree on a concrete answer to this simple question, how can we design autonomous systems for complex situations?

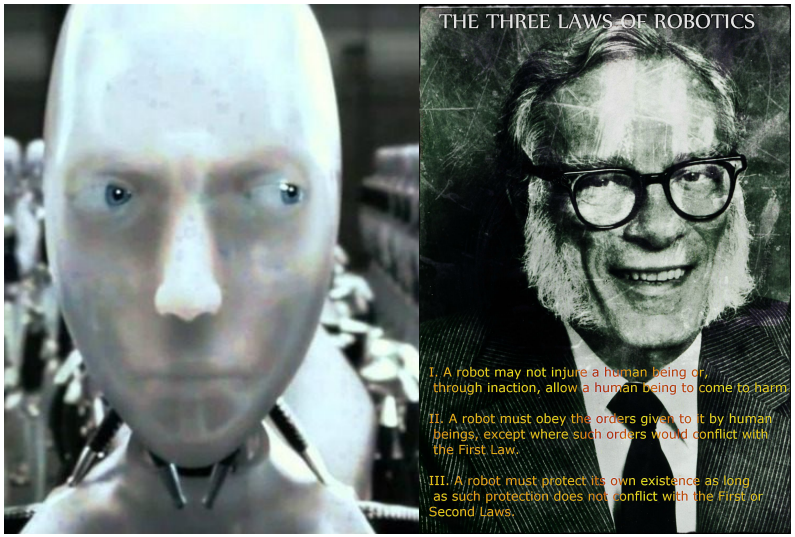
1. **Equal intentional rescues:** The algorithm could randomly with equal probability select one of the three persons not to endanger and act accordingly. Each person is intentionally not endangered in 1/3 of the cases. However, the postdoc distribution will be biased to not endanger the second and third person in 2/3 of the cases, as they stand together. The average number of casualties is 4/3 ~ 1.33 per dilemma situation.
2. **Equal post-hoc rescues:** The algorithm could randomly with equal probability select one side, irrespective of the number of people present, to not endanger. The resulting distribution is that all subjects are not endangered in 50% of the cases with an average number of casualties of 3/2 = 1.50 per dilemma situation.
3. **Minimal casualties:** The algorithm could always chose not to endanger the larger group. The resulting distribution is completely biased, subject 1 is always endangered and subjects 2 and 3 are never endangered. The average number of casualties results in 1.00 per dilemma situation.

[your choice]

HOW TO SECURE SECURE OPERATION? #1

(A) Testing. However, we have to consider e.g. the halting problem, adversarial attacks and complexity of real-time systems. This will not be sufficient.

(B) 3 root laws. How to implement?



Salge & Polani (2017) Front Robot AI

Empowerment: Formal Definition

$$\mathfrak{E}^{(k)}(w_{t-k}) := \max_{p(\mathbf{a}_{t-k}^{(k)}|w_{t-k})} I(\mathbf{A}_{t-k}^{(k)}; S_t|w_{t-k})$$

(C) Empowerment. To be in command of your own fate has positive value (Jung et al., 2011). This can be mathematically formalised as empowerment (Klyubin et al., 2005; 2008; Nehaniv et al., 2007). Importantly, this does not make assumptions on the direct function of reward. For example, food is good, as it allows you to live and act.

Autonomous systems need a prefrontal cortex, aka. ethics by design.

HOW TO SECURE SECURE OPERATION? #2

We can not at the same time not discriminate and differentiate!?

"#9 Bei unausweichlichen Unfallsituationen ist jede Qualifizierung nach persönlichen Merkmalen (Alter, Geschlecht, körperliche oder geistige Konstitution) strikt untersagt. … "

Autonomous cars (should) do what we ask them to do.

We must be honest to ourselves.

Summary & acknowledgements

- Kaspar K, Krapp V
- Sütfeld LR, Gast R, Pipa G
- Faulhaber AK, Dittmer A, Blind F, Wächter MA, Timm S, Sütfeld LR, Stephan A, Pipa G
- Bergmann LT, Schlicht L, Meixner C, Pipa G, Boshammer S, Stephan A
- Kallioinen N, Pershina M, Ruda I, Mukhina A, Nezami FN, Almstedt JP, Edelkott L, Finger D, Ghazaryan G, Zeiser J
- Oliveira AS, Schlink BR, Hairston WD, Ferris DP
- Kakerbeck K, König SU
- Nezami FN, Wächter MA

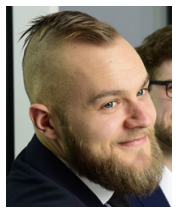

Bergmann L


Nezami FN


Stefan A


Sütfeld LR


Pipa G


Wächter M

- Introducing autonomous cars promises huge benefits..
- With respect to autonomous cars the results make clear suggestions. (1) Human decisions in a VR trolley dilemma can be very well described by a value of life concept. (2) The perspective has only a small modulatory effect. (3)The acceptability of actions by human drivers and autonomous cars is largely congruent. All this is good news for engineers and allows to optimise a clear objective.
- However, empirically not all human avatars are treated equally and quantitative comparisons dominate behaviour. Further, the perspective has a modulatory effect, most notably in the context of a sidewalk. This has to be subject of a public discussion. Maybe robots have to be the better humans, or we revise some of the ethics.
- Finally, these questions beg for an answer of the fundamental problem of testing/validating autonomous systems. Can we achieve ethics by design?

erc DFG CaN CTA SEVENTH FRAMEWORK PROGRAMME Europa fördert Niedersachsen Universität Osnabrück AKADEMIE DER WISSENSCHAFTEN IN HAMBURG