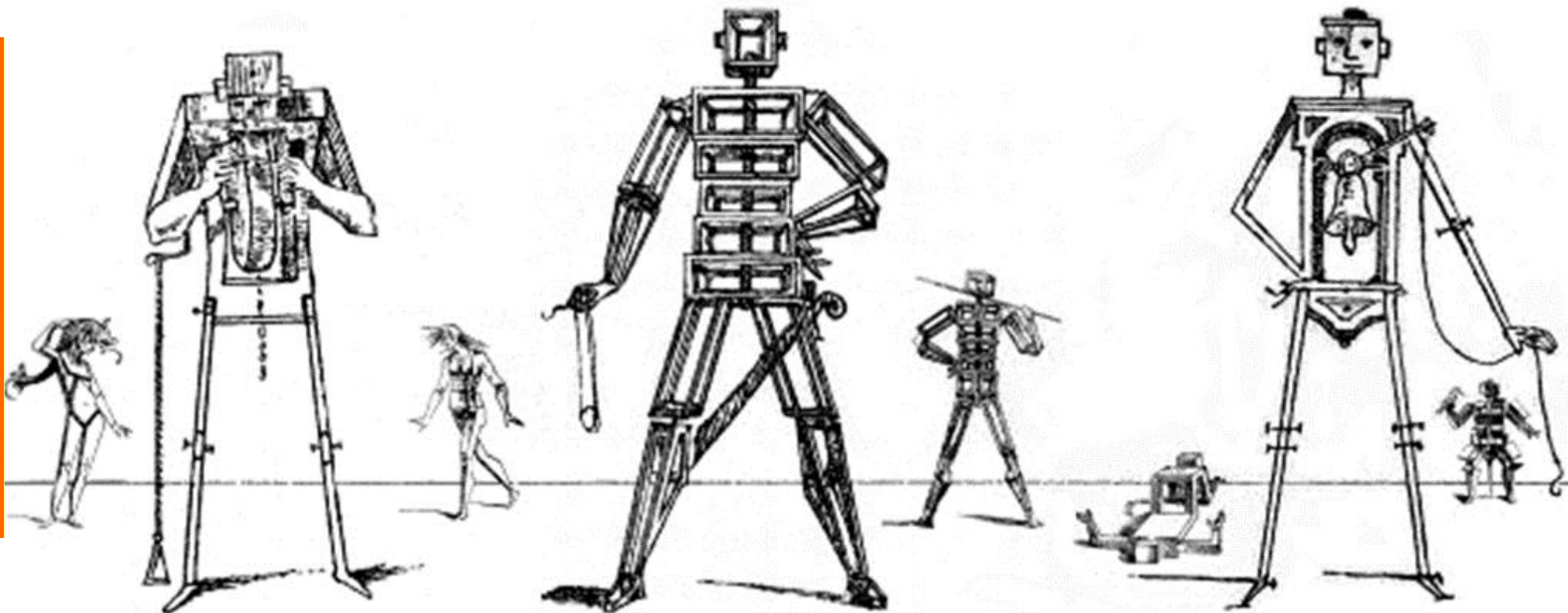


# Towards a machine ethics

Prof. Dr. Oliver Bendel

With contributions by Dr. Gwendolin Wilke

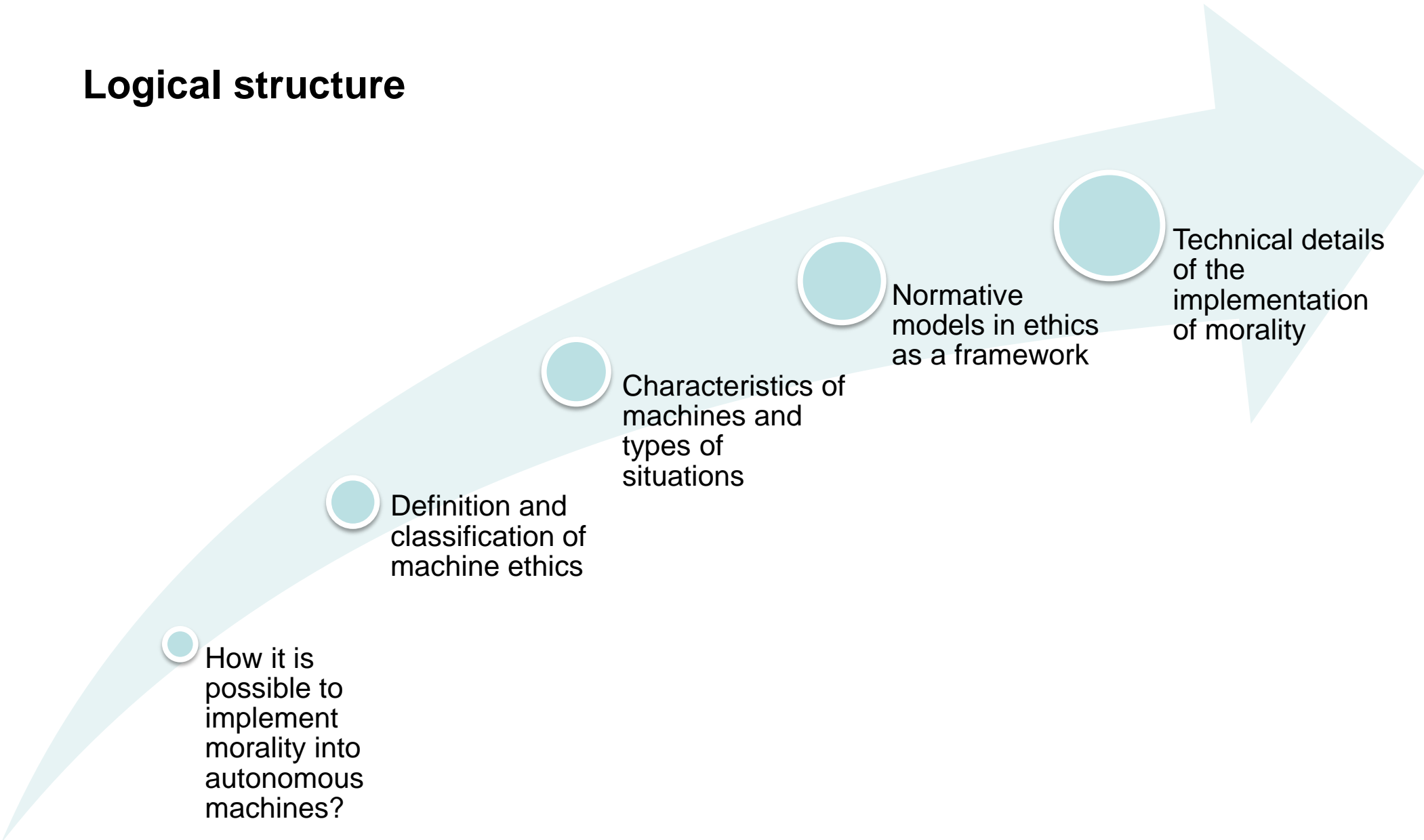


## Introduction

In this presentation the young field of **machine ethics** is explored. Technology assessment (TA) is concerned with the **consequences of technical developments**, and some of its topics have **moral dimensions**. It could be valuable for TA to keep an eye on machine ethics.

The main question of this presentation is **whether and how it is possible to implement morality into autonomous machines**. The answer (or the attempt of an answer) is based on the review of existing literature, own classifications, considerations and derivatives.

## Logical structure



How it is possible to implement morality into autonomous machines?

Definition and classification of machine ethics

Characteristics of machines and types of situations

Normative models in ethics as a framework

Technical details of the implementation of morality

## Term and classification of machine ethics

**Machine ethics** pays attention to the morality of **autonomous machines** such as agents, chatbots, algorithmic trading computers, robots of different stripes, and unmanned ground or air vehicles.

It can be seen as a part of **information ethics** and **technology ethics** (cf. Bendel 2012b). From this point of view, it is only another field of applied ethics.

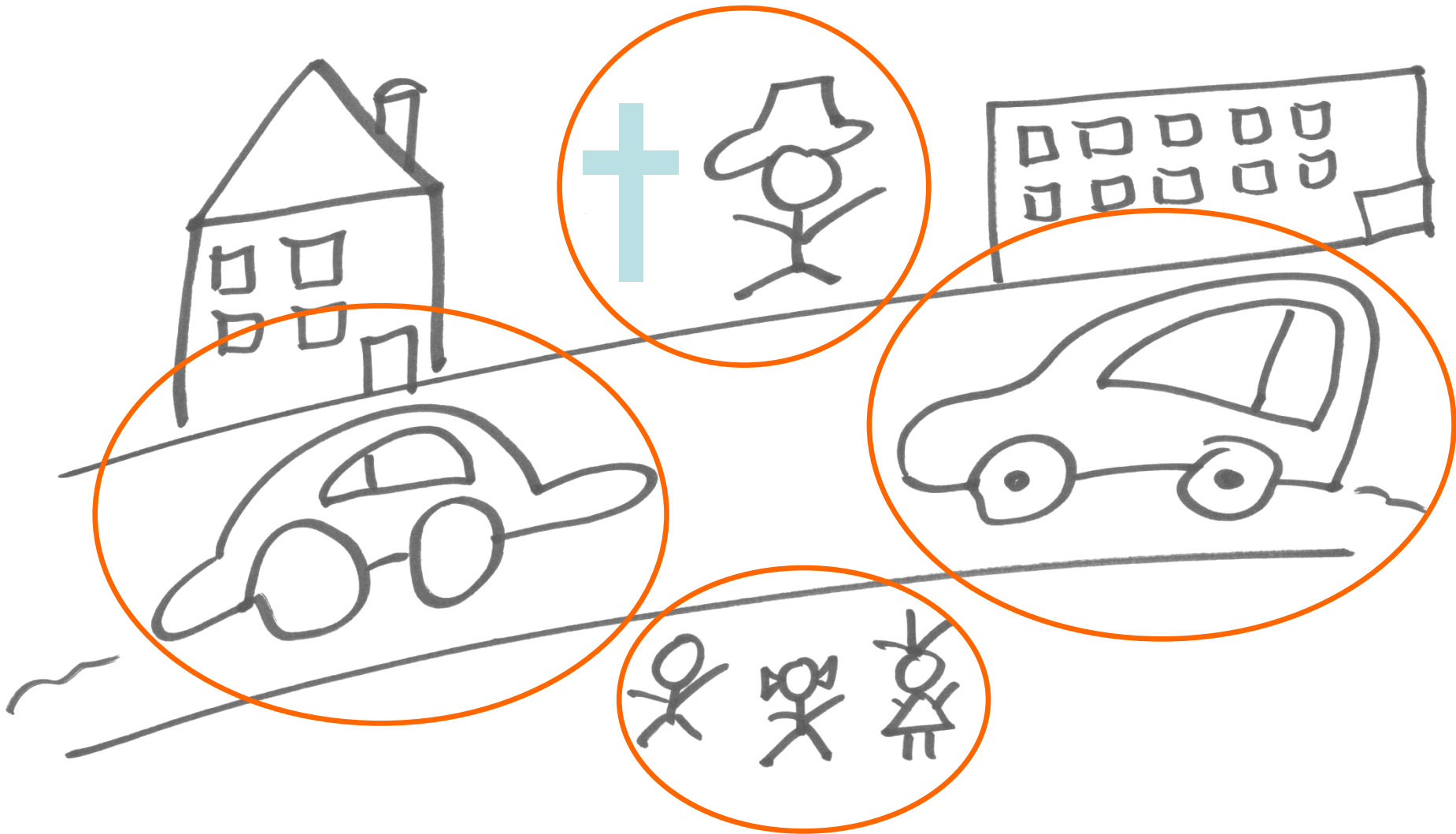
But with good reason, it can also be understood as a **counterpart of human ethics** (cf. Bendel 2012e). From this perspective, machine ethics is a new form of ethics.

## Abbreviated literature review

For this presentation, two books were evaluated, “**Machine Ethics**” by Michael and Susan Leigh Anderson as editors (2011) and “**Robot Ethics**” by Patrick Lin, Keith Abney and George A. Bekey as editors (2012):

- Some authors refer to Isaac Asimov and his **Three Laws of Robotics** and reflect upon the **basic meanings and implications** of machine ethics (cf. Clarke 2011).
- Some authors discuss **deontological** or **teleological normative models** with respect to the use for machine morality.
- James Gips focuses on **virtue ethics** (cf. Gips 2011).
- Bruce M. McLaren promotes a **case-based reasoning** (cf. McLaren 2011), and Marcello Guarini gives a **neural network approach** (cf. Guarini 2011).

## The story of NAC



## Characteristics of machines

We can distinguish between different **characteristics of machines**:

- Machines have different **tasks** and **fields of use**.
- Some are only **partially autonomous** (acting under human command) while others are **completely autonomous** within their area of action.
- Most systems **act and decide**, some are able to **show emotions**, and some are able to **communicate in a natural language**.

## Types of situations

Furthermore, it is useful to distinguish between various **types of situations** in which machines act.

- We must identify their **content** and their **coordinates**, perhaps also their **cultural, political, and legal contexts**.
- We have to draw a distinction between situations in which machines must **act fast** or **not so fast** and in which **things, animals, or people** are affected.
- There is also a difference between **closed situations** and **open situations**, between **simple situations** and **complex situations**, as well as between situations in the **present** and in the **future**.



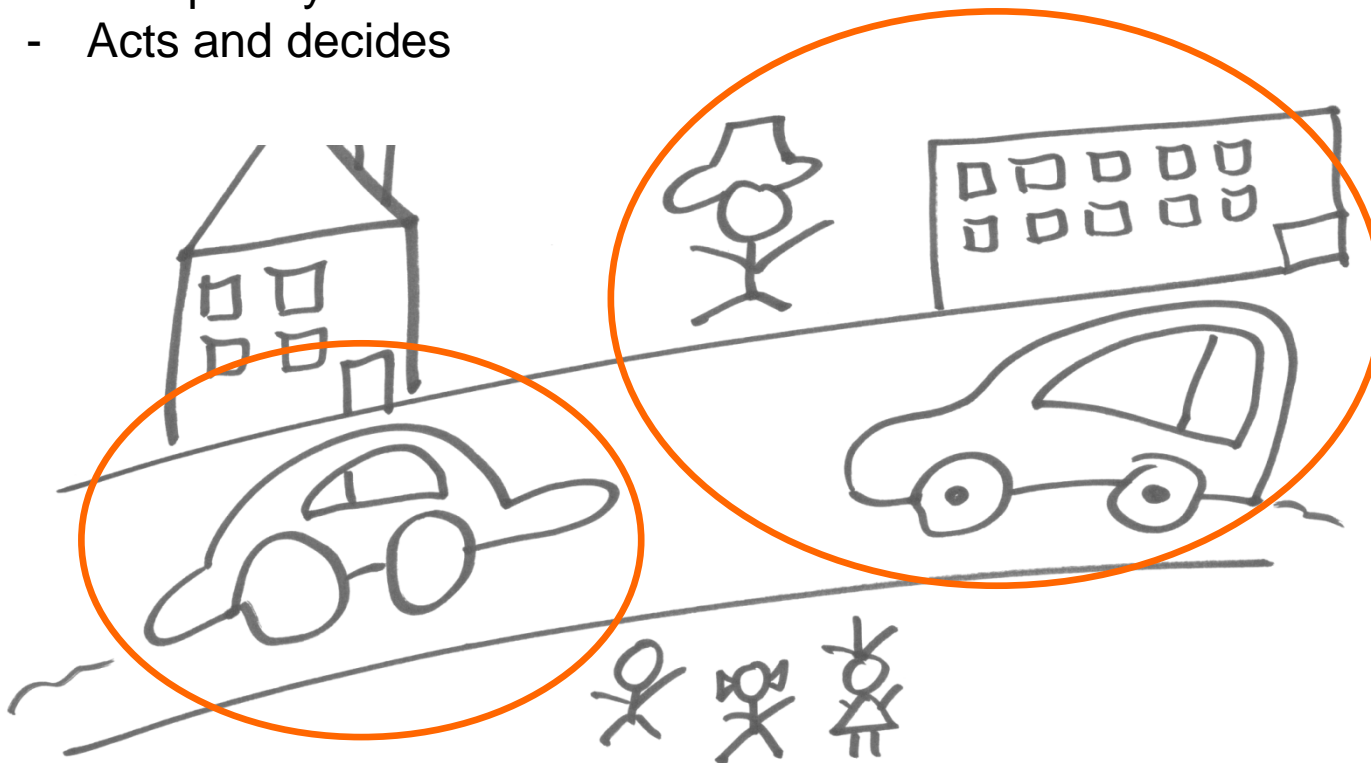
## Description of NAC

Characteristics of machine:

- Drives on the road
- Completely autonomous
- Acts and decides

Types of situation:

- Accident on the road during the day
- City in Europe
- Machine must act fast
- Things and people are affected
- Open situation
- Complex situation
- Situation in the present



## Normative models for machine ethics

There exist a number of **normative models** in ethics to assess and justify morality. According to Pieper (cf. Pieper 2007), seven fundamental models can be distinguished:

1. The **transcendental model** in the form of **deontological ethics** (Immanuel Kant) seems promising. Machines can carry out their duties, if these are formulated as rules. Although rules can clash, and situations can be so complex that it is not clear which rule should be chosen.
2. The **existential approach** (Sören Kierkegaard, Jean-Paul Sartre) perhaps does not make sense for the machine ethics, because the existence of human beings is addressed. A transfer to machine ethics would be difficult, especially as this approach is concerned only with the unit of human qualities.

## Normative models for machine ethics

3. We should think about the **eudemonistic model** (Aristotle) in the form of **teleological ethics**. Machines are able to evaluate consequences. And they can observe their environment and build their own database. Of course, it is generally not easy to look into the future.
4. The **contracting theory** (John Rawls) is also a promising approach, but applies to the behavior of machines amongst themselves. Some autonomous systems can communicate and reach an agreement. It is important that their decisions are not addressed against us and fit within our idea of justice.
5. In the **traditional framework, virtue ethics** (also Aristotle) plays an important rule. It may be curious to award virtue to a machine. However, the ability to self-learn could be a starting point for a kind of virtue or conscience.

## Normative models for machine ethics

6. In the case of the **materialistic approach** (materialists and Marxists) the difficulties are as large as in the case of existentialism. If matter is to be the basis for mentality and morality, it has only to be arranged in the right way. The question remains: how it can be arranged?
7. The **life-world model** (Wilhelm Schmid, Wilhelm Vossenkuhl, and Otfried Höffe) is an interesting candidate because of its eclectic character. With a combination of different approaches, the systems could overcome their limitations in the context of ethics. But what is the right mix?

## Deontological model

In the deontological model, **duties** are the point of departure. Duties can be translated into **rules**. It can be distinguished between (simple) **rules** and **meta rules**.

A machine can follow simple **rules**. Rule-based systems can be implemented as **formal systems** (also referred to as axiomatic systems), and in the case of machine ethics, a set of rules is used to determine which actions are morally allowable and which are not.

Since it is not possible to cover every situation by a rule, an inference engine is used to **deduce new rules** (respectively recommendations) from a small set of simple rules (called axioms) by combining them. The morality of a machine is comprised of the set of rules that are deducible from the axioms.

## Deontological model

The disadvantage of using formal systems is that many of them work only in **closed worlds** like computer games. What is not known is assumed to be false. This is in drastic conflict with real world situations, where rules can conflict and it is impossible to know everything about the environment. Here, a **prioritization of rules** can be provided in order to restore consistency, like in the Three Laws of Robotics from Asimov, and **meta rules** can be useful to evaluate them. Another approach to avoid a closed world assumption is to utilize **self-learning algorithms**, such as case-based reasoning approaches.

## Teleological model

In the context of the **teleological model**, the **consequences of an action** are assessed. The machine must know the consequences of an action and what the action's consequences mean for humans, for animals, for things in the environment, and, finally, for the machine itself.

The system must also be able to assess whether these consequences are **good or bad**, or if they are acceptable or not, and this assessment is not absolute, of course.

An implementation approach that allows for the consideration of potentially contradictory subjective interests may be realized by **decentralized reasoning approaches** such as agent based systems. In contrast to this, **centralized approaches** may be used to assess the overall consequences for all involved parties (a formula, for example).

## Teleological model

In the teleological model, it is essential that a machine is able to address not only present facts, but also possible future states of the world in order to allow for the assessment of an action's consequences. Therefore an implementation of morality must provide **prospective abilities**. When we refer to something in the future, it may be uncertain or vague (cf. Papaioannou 2013).

When implementing a moral framework for machines, the inherently imperfect knowledge about the future can be dealt with by calculi of imperfections such as **fuzzy logic**, **possibility theory** or **probability theory**.



## Teleological model



## Traditional model

In the context of machine ethics, the **traditional model** may mean that a machine needs to acquire **virtues** such as wisdom, justice, courage and temperance, and develops a character which includes a set of them. The “morally right” action implicitly **follows from this character**.

Similar to the rule-based approach, its virtues may be **prioritized** or formed in a special way in order to adjust them to the intended character of the machine. Another more flexible approach is to use **adaptive or self-learning systems** such as machines with genetic algorithms, agent-based systems or neural networks.

## Additions and combinations

Last but not least, human beings may act as **reference persons**, and **social media** may serve as a moral input – rather questionable possibilities which were stated in the short story about NAC (cf. Bendel 2012b).

Perhaps a combination of all these approaches will be successful, e.g. in the **life-world model**, though the name seems to be curious in this context, because there is no real life of a machine.

Maybe there is a need to leave the classical approaches behind and to develop a new model that **suits both human beings and machines**.

## Summary and outlook

The speaker is **sceptical about the possibility of implementing a moral code** in a machine in a satisfactory manner. Moreover, the requirements of machine processing could be different from system to system (and even from situation to situation), and an approach which works well in one environment can fail in another.

However, there will be **substantial interest** from the fields of **industry** and **military** that would like to bring their solutions into the market respectively to the areas of conflict, and, in a different sense, of **philosophy** to solve some of the central questions. To say it from the philosophical point of view: Machine ethics will be the **touchstone of ethics in general**.

## Summary and outlook

Will it be also a **touchstone of technology assessment**? Yes, in the sense that TA has to integrate new fields. And in the sense that it has to ask fundamental questions now:

- How should technology be in the future?
- Do we want to have autonomous systems at all?
- Do we want to have machines that think and feel?
- That behave morally, as subjects of morality, and that are even objects of morality some day?

Perhaps **technology assessment will be totally different in the future**  
– **because technology will be totally different.**

## Literature

(Abney 2012) Abney, Keith. Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed. In: Lin, Patrick; Abney, Keith; Bekey, George A. (Ed.). Robot Ethics: The Ethical and Social Implications of Robotics. The Mit Press, Cambridge, MA 2012, pp. 35–54.

(Anderson/Anderson 2011) Anderson, Michael; Anderson, Susan Leigh (eds.). Machine Ethics. Cambridge University Press, Cambridge 2011.

(Bekey 2012) Bekey, George A. Current Trends in Robotics. In: Lin, Patrick; Abney, Keith; Bekey, George A. (ed.). Robot Ethics: The Ethical and Social Implications of Robotics. The Mit Press, Cambridge, MA 2012, pp. 17–34.

(Bendel 2012a) Bendel, Oliver. Die Medizinethik in der Informationsgesellschaft: Überlegungen zur Stellung der Informationsethik. In: Informatik-Spektrum, November 2012 ("Online-First"-Artikel auf SpringerLink).

(Bendel 2012b) Bendel, Oliver. Die Moral der Maschinen: Überlegungen zur Maschinenethik. In: inside-it.ch, 24 October 2012. Via <http://www.inside-it.ch/articles/30517>.

(Bendel 2012c) Bendel, Oliver. Die Medizin in der Moral der Informationsgesellschaft: Zum Verhältnis von Medizinethik und Informationsethik. In: IT for Health, 3 (2012) 2. pp. 17–18.

(Bendel 2012d) Bendel, Oliver. Informationsethik. In: Gabler Wirtschaftslexikon. Gabler/Springer, Wiesbaden 2012. Via <http://wirtschaftslexikon.gabler.de>.

(Bendel 2012e) Bendel, Oliver. Maschinenethik. In: Gabler Wirtschaftslexikon. Gabler/Springer, Wiesbaden 2012. Via <http://wirtschaftslexikon.gabler.de>.

(Clarke 2011) Clarke, James. Asimov's Laws of Robotics: Implications for Information Technology. In: Anderson, Michael; Anderson, Susan Leigh (eds.). Machine Ethics. Cambridge University Press, Cambridge 2011. pp. 254–284.

(Edgar 1997) Edgar, Stacey L. Morality and Machines: Perspectives on Computer Ethics. Jones and Bartlett Publishers, Sudbury 1997.

## Literature

(Egger 2013) Egger, Lenhard. The morality of avatars. Term paper for the course “ToBIT” at the School of Business of the University of Applied Sciences Northwestern Switzerland (FHNW). Grey literature. Olten 2013.

(Guarini 2011) Guarini, Marcello. Computational Neural Modeling and the Philosophy of Ethics: Reflections on the Particularism-Generalism Debate. In: Anderson, Michael; Anderson, Susan Leigh (eds.). Machine Ethics. Cambridge University Press, Cambridge 2011. pp. 244–253.

(Gips 2011) Gips, James. Towards the Ethical Robot. In: Anderson, Michael; Anderson, Susan Leigh (eds.). Machine Ethics. Cambridge University Press, Cambridge 2011. pp. 316–334.

(Hänßler 2011) Hänßler, Boris. Schaltkreise für Schuld und Moral. In: derFreitag, 14 June 2011. Via <http://www.freitag.de/wissen/1124-schaltkreise-f-r-schuld-und-moral>.

(Höffe 2008) Höffe, Otfried. Lexikon der Ethik. 7., neubearb. und erweit. Auflage. München: C. H. Beck 2008.

(Lin 2012) Lin, Patrick; Abney, Keith; Bekey, George A. (Ed.). Robot Ethics: The Ethical and Social Implications of Robotics. The MIT Press, Cambridge, MA 2012.

(McLaren 2011) McLaren, Bruce M. Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions. In: Anderson, Michael; Anderson, Susan Leigh (eds.). Machine Ethics. Cambridge University Press, Cambridge 2011. pp. 297–315.

(Papaioannou 2013) Papaioannou, Yiouli. Fuzzylogik in Wissenschaft, Recht und Ethik. In: faz.net, 26 January 2013. <http://www.faz.net/aktuell/wissen/atomium-culture/logik-mit-unschaerfen-fuzzylogik-in-wissenschaft-recht-und-ethik-12029866.html>.

(Pieper 2007) Pieper, Annemarie. Einführung in die Ethik. 6. revised and updated edition. A. Francke Verlag, Tübingen and Basel 2007.

(Schnyder 2013) Schnyder, Matthias. Towards a Machine Ethics. Term paper for the course “ToBIT” at the School of Business of the University of Applied Sciences Northwestern Switzerland (FHNW). Grey literature. Olten 2013.

(Trenkamp 2012) Trenkamp, Oliver. Bettina Wulff gegen Google: Attacke auf den Algorithmus. In: Spiegel Online, 8. September 2012. Via <http://www.spiegel.de/netzwelt/web/bettina-wulff-klagt-gegen-google-a-854710.html>.