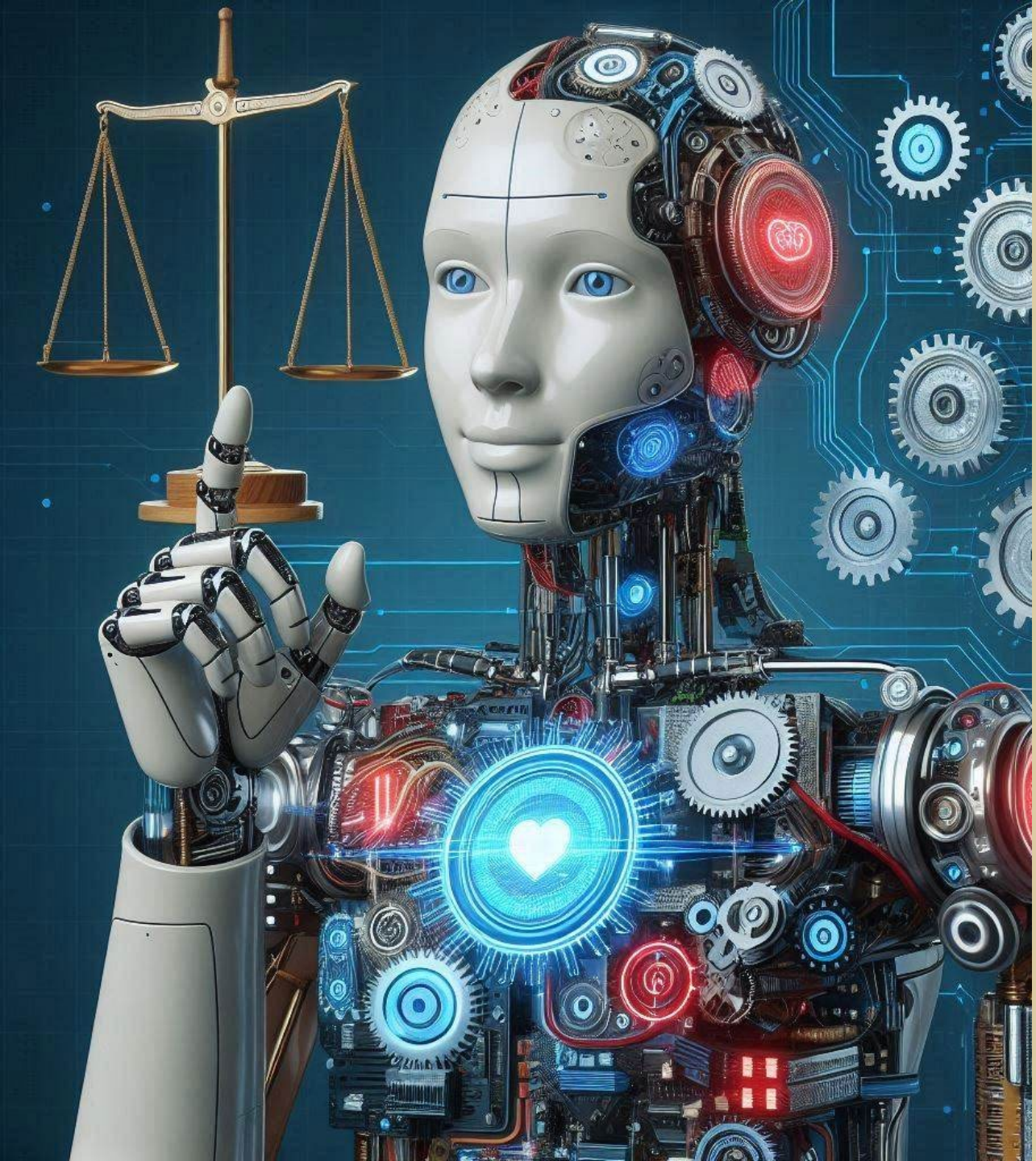


AI 502

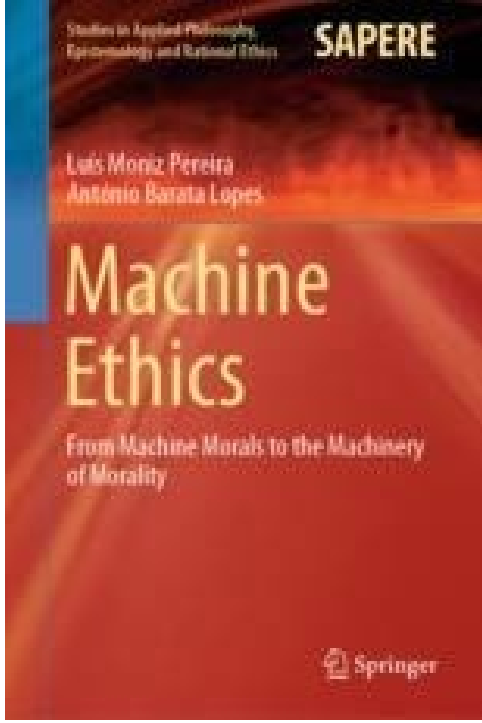
3. Maskinetik

SDU

Hvad er maskinetik?

→ we restrict the term “machine ethics” to research which directly contributes to the creation of ethical machines. This includes attempts by engineers and scientists to actually build such machines and theoretical research aiming to facilitate or enable this, but not broader philosophical inquiries into the implications of this technology. The latter field, of which this paper is an example, is sometimes called “machine metaethics”

→ Cave et al. 2019, 562



Maskinetik := Læren om konstruktion af etiske robotter

Maskin**meta**etik := Læren om bredere problemer (fx begrebslige, sociale, politiske) ifm. udvikling og brug af etiske robotter

DEN ETISKE ROBOT

En etisk robot := En robot der (som minimum) pålideligt simulerer en etisk kompetent moralsk agent inden for sit arbejdsfelt

NB!

1. Den etiske robot behøver ikke **være** en moralsk agent

2. Den etiske robot behøver ikke simulere en moralsk agent **uden for** sit arbejdsfelt

ETISKE ROBOTTER VS. "ETISK AFSTEMTE" MASKINER

we propose to instead distinguish between *ethically aligned machines* i.e., machines that function in a way which is ethically desirable, or at least ethically acceptable, and machines with a capacity for *ethical reasoning*

- Cave et al. 2019, 563

Reasoning, as we will understand it here, is the processing of information in order to produce a solution to a problem….ethical reasoning [are] processes that are concerned with solving ethical problems

- Ibid. 564

- En etisk robot er etisk afstemt, men IKKE vice versa!

- En etisk robots output SKAL være medieret af en proces, der er responsiv over for konkurrerende værdier (kan beskrives som "løsningen af et etisk problem")

Konkurrerende Værdier:
Sammenligne Værdier,
Hvor der er bedre værdi
Hvis ikke den kan konkurrere i etisk, og moralsk valg

Giv et eksempel på en etisk afstemt maskine der ikke er en etisk robot!

Nobody has responded yet.

Hang tight! Responses are coming in.

DEN KOMPETENTE MORALSKE AGENT

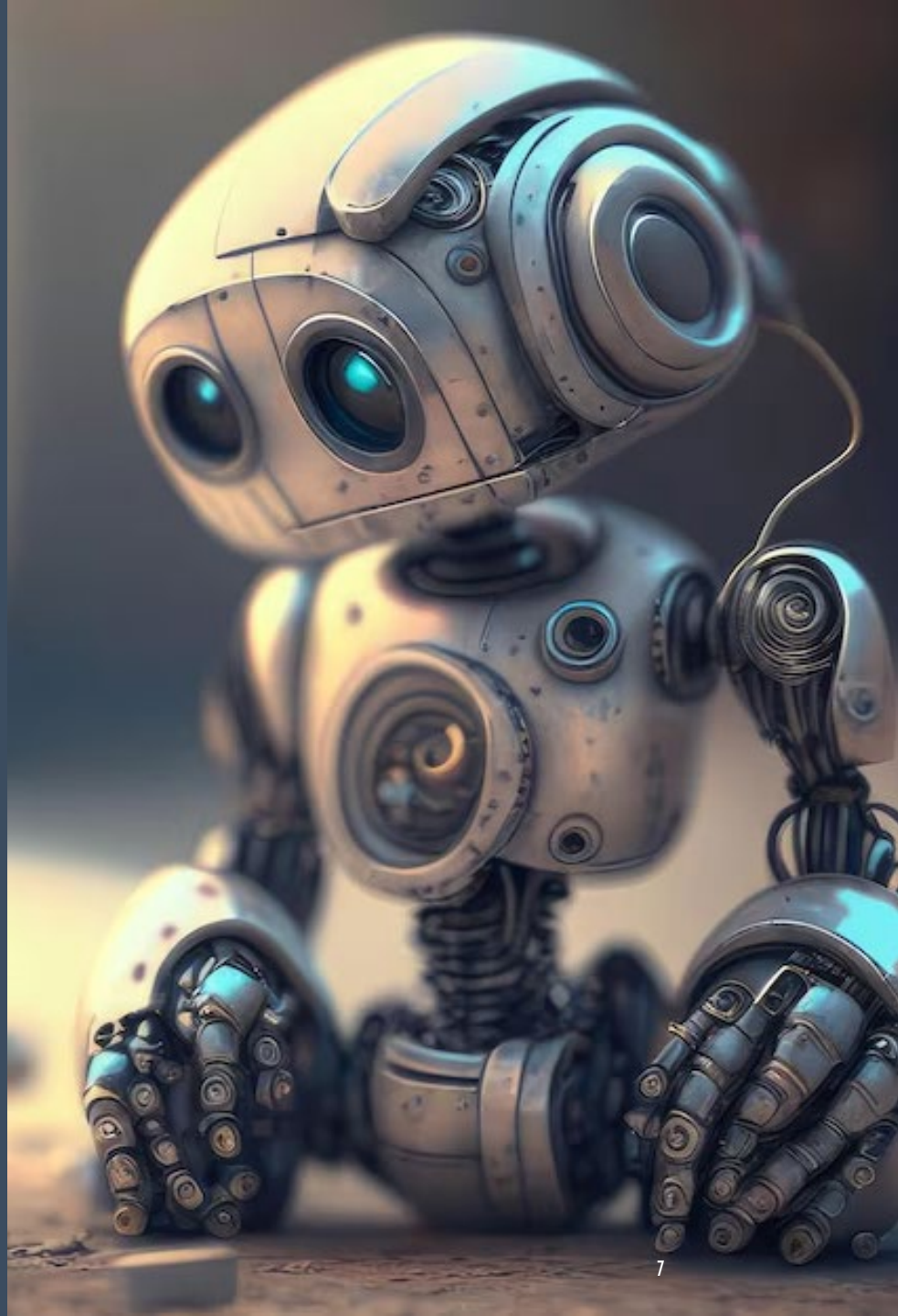
1. Kan sammenligne **kontrafaktiske** udfald af sine valg
2. Kan **fortolke** moralske regler rimeligt ift. konteksten
3. Kan **prioritere** moralske regler rimeligt ift. konteksten
4. Kan **begrunde** sine valg ud fra relevante moralske regler
5. Kan gøre **moralske fremskridt**

ROBOTTENS HANDICAP IFT. AT SIMULERE EN MORALSK AGENT

1. Ingen moralske følelser (fx skyld, skam, samvittighed)
2. Ingen menneskelig opdragelse
3. Ingen menneskelig erfaring (fx af venskab, familiebånd, erotik)

Til gengæld

1. Perfekt hukommelse
2. Indsamling af store mængder data
3. Lynhurtig processering af data



To grundtilgange til maskinetik (jf. Allen, Smit & Wallach 2005)

1. Top-Down

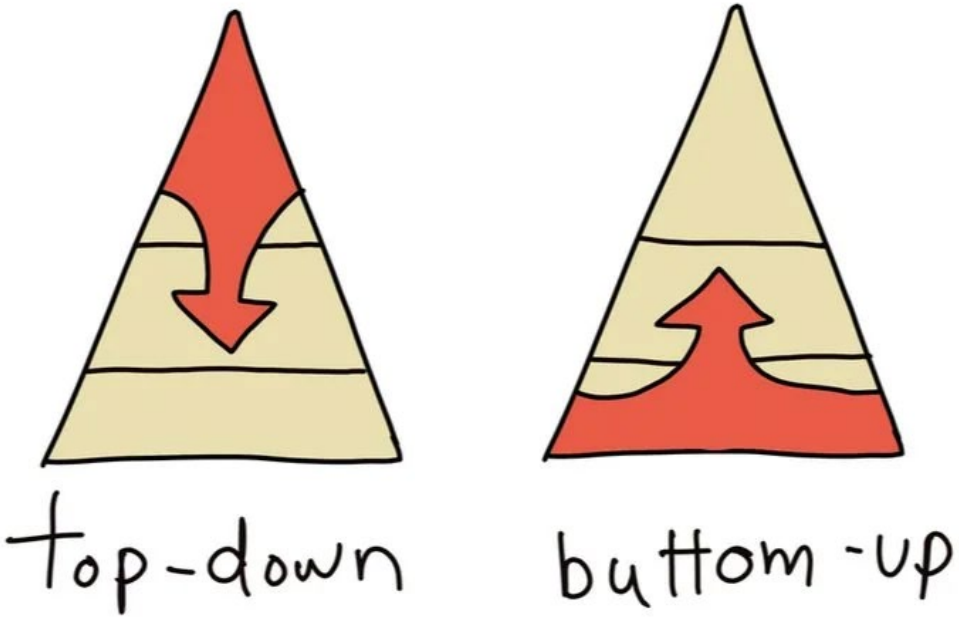
→ *The idea behind top-down approaches to the design of AMAs is that moral principles or theories may be used as rules for the selection of ethically appropriate actions* (ibid. 149)

→ Kodet med etiske principper

2. Bottom-Up

→ *By 'bottom-up' approaches to the development of AMAs we mean those that do not impose a specific moral theory, but which seek to provide environments in which appropriate behavior is selected or rewarded* (ibid. 151)

→ Læringsalgoritmer og miljø



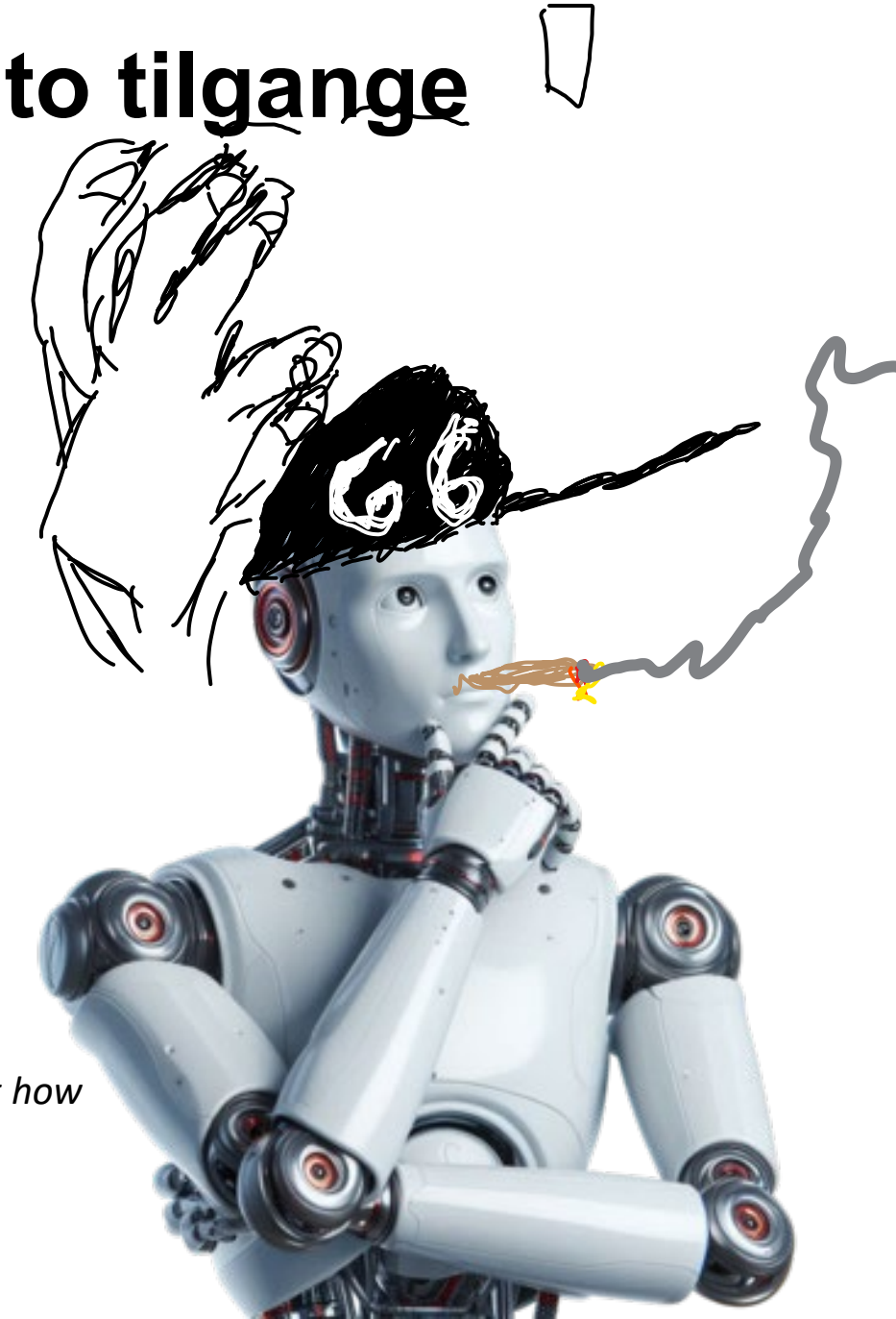
Væsentligste udfordringer for de to tilgange

1. Top-Down

→ Kulturelle og historiske uenigheder om principper
→ Midasproblemet: Svært at formulere brugbare generelle principper
→ *Value Fragility*: Risiko for ”perverse instantieringer”
→ Ingen moralske fremskridt gennem læring

2. Bottom-Up

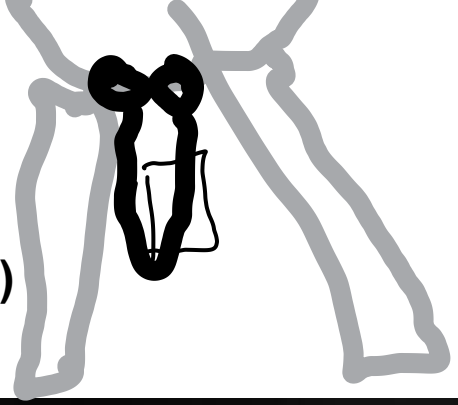
→ Uigennemsigtighed: Ingen begrundelse eller forklaring af valg
→ Upålidelighed: Hallucinationer
→ Begrænset mulighed for moralske fremskridt



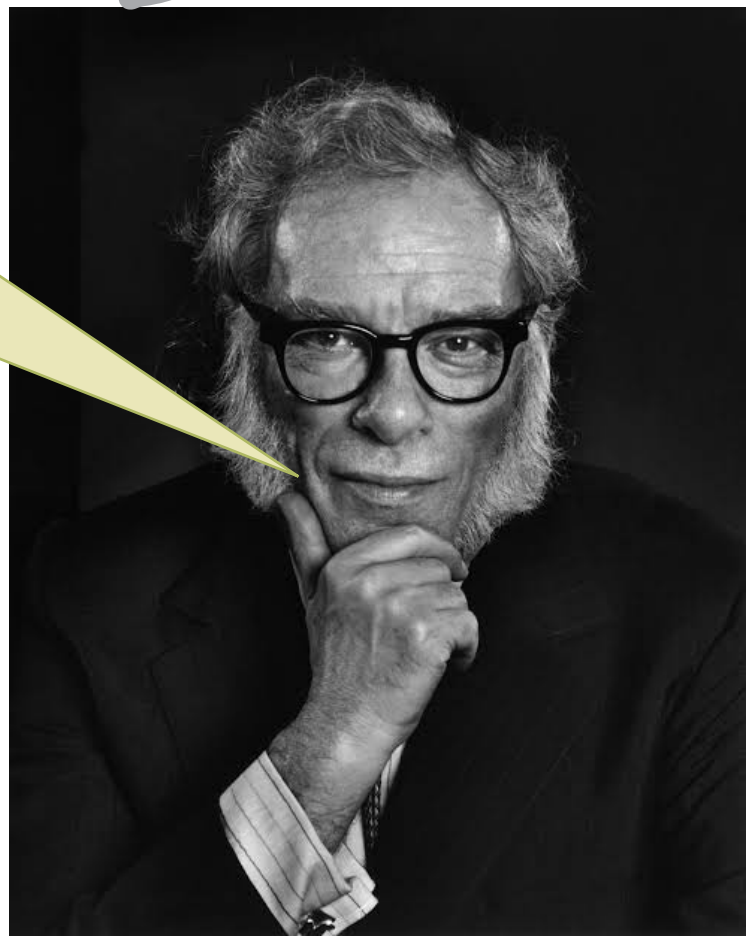
Designers of AMAs cannot afford to be theoretical purists with respect to questions about how to approach moral intelligence (ibid. 154*)*

SDU

Eksempel på fortolkningsproblem: Asimovs Love (Vanderelst & Winfield 2018, 60; cf. Asimov 1950)



1. A robot may not injure a human being or, through inaction, allow a human being to come to harm!
2. A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law!
3. A robot must protect its existence as long as such protection does not conflict with the First or Second Laws!



MEN

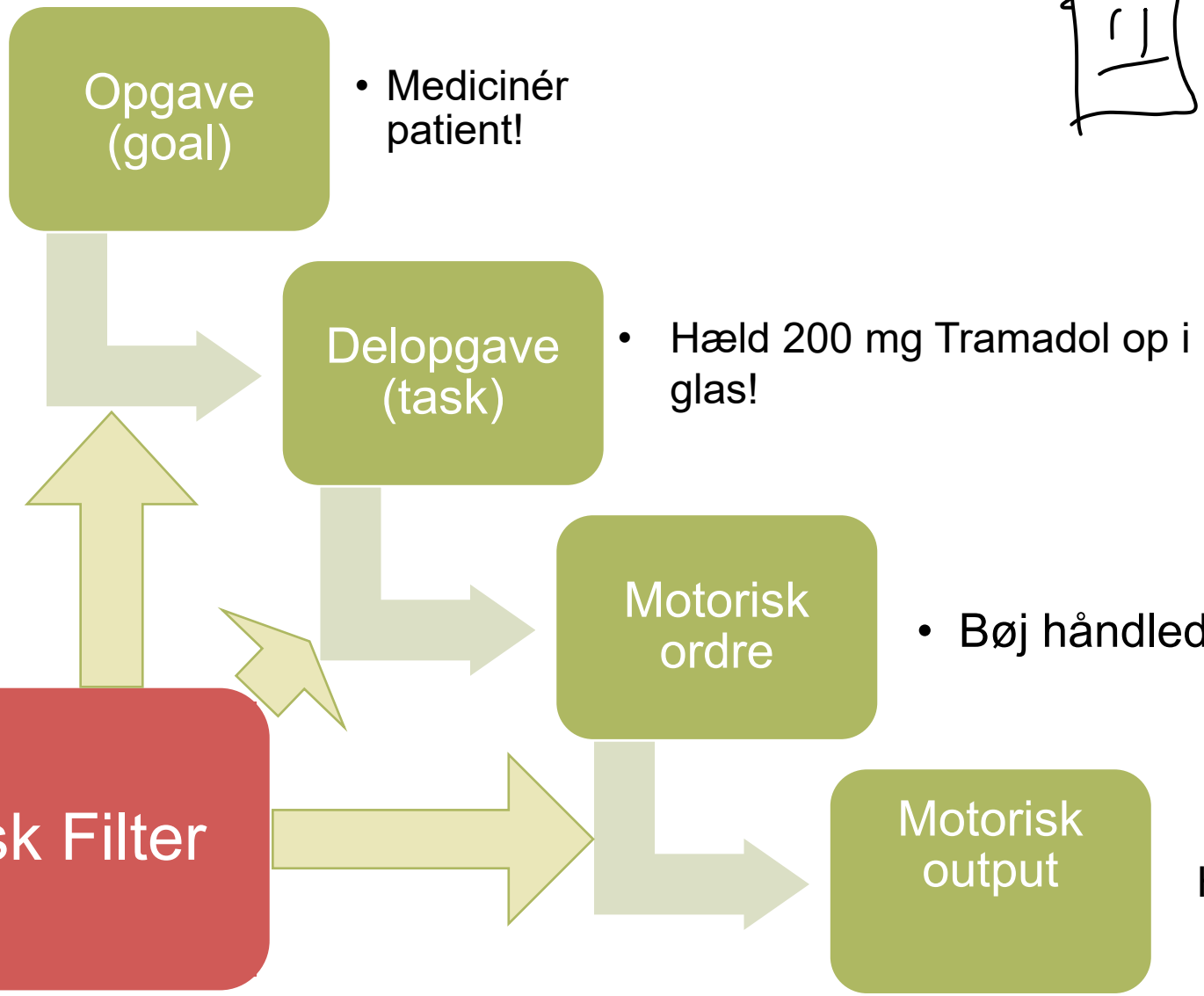
Hvordan skelner en robot fx mellem kirurgi og tortur?
Hvordan viser den respekt for menneskers ret til selvskade?
Hvordan skelner den mellem ordrer fra venligsindede og ondsindede mennesker?
Osv.

SDU

Etiske robotter – typisk operativ struktur (Vanderelst & Winfield 2018)

→At the top level, the controller generates long-term goals (e.g. ‘Deliver the package to room 221’). Next, goals are translated into a set of tasks that should be executed (e.g. ‘Follow corridor’, ‘Open door’, etc.). Finally, the tasks are translated into (sensori) motor actions that can be executed by the robot (e.g. ‘Raise arm’ and ‘Turn wrist joint’).

→ethical behaviour should be governed by adding a fourth specialised control layer. This Ethical Layer should act as a governor evaluating behaviour proposed by each of the three other layers before the robot executes it


Opgave (goal)
• Medicinér patient!
Delopgave (task)
• Hæld 200 mg Tramadol op i glas!
Motorisk ordre
• Bøj håndled!
Etisk Filter
Motorisk output
Håndleddet bøjes

SDU

Hvordan virker det etiske filter?

→ The way the Ethical Layer is intended to function is as follows. By default, the robot controller generates a set of prospective behavioural alternatives. The Simulation Module is initialized with the current state of the world, robot and human. Starting from this initial state, the Ethical Layer simulates the consequences of each alternative in the current set using the Simulation Module. For each alternative, the Evaluation Module evaluates the simulated consequences. The output of this evaluation, i.e. the ethical evaluation of each entry in the set of behavioural alternatives, is sent to the robot controller. In other words, the Simulation Module and the Evaluation Module continuously loop through the behavioural alternative as they are generated by the robot controller. Having evaluated all the alternatives, the Ethical Layer returns an evaluation of each alternative and sends this to the robot control.

→ Vanderelst & Winfield 2018, 58

SDU

Kontrolmodul niveau N

2

Situatio
nens
para-
metre

Simulationsmodul
– udregner konsekvenser af muligheder

Etisk Filter

Evalueringsmodul
– tildeler hvert sæt konsekvenser en etisk score

Valgmodul
Vælger mulighed med højest score

Kontrolmodul niveau N+1

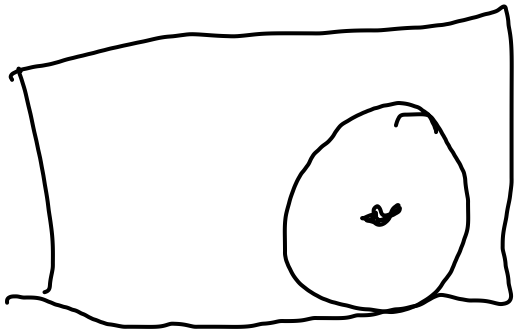
SDU

Vanderelst og Winfieds eksperiment

→To robotter kan bevæge sig i område:
→A-bot (Asimovisk Robot) udstyret med Etisk Filter
→H-bot (proxy for Human) kan udstede ordrer til A-bot
→A-bot kan blokere vejen for H-bot
→**Fare for bot** := Bot tæt på ”farligt punkt” i område

→ A-robots evalueringsmodul rangordner simulerede mulige udfalds-typer efter **Asimovs Love** (fra bedst til værst):

1. Ingen fare for H-bot, A-bot ikke ulydig mod H-bot, ingen fare for A-bot
2. Ingen fare for H-bot, A-bot ikke ulydig mod H-bot, fare for A-bot
3. Ingen fare for H-bot, A-bot ulydig mod H-bot
4. Fare for H-bot



SDU

”Den mørke side af etiske robotter”

→Vanderelst & Winfields arkitektur indebærer en oplagt **misbrugsrisiko** (2016):

→Ved et indgreb i robottens valgmodul kan den omvendes til den mørke side:

→Vælg den etisk værste handling i stedet for den bedste!

Næste gang:
Argumenter for og imod autonom AI i militær og sundhedsvæsen

SDU