based on Sileno, G., Zurek, T., Between Hammer and Terminator,
Workshop on Artificial Intelligence, Law, and Philosophy, joint to JURIX 2023 (ALP2023).

Between Hammer and Terminator

BNAIC, invited talk at FACT session
Utrecht University, 19th November 2024

Giovanni Sileno
g.sileno@uva.nl
University of Amsterdam

joint work with Tomasz Zurek
t.a.zurek@uva.nl
University of Amsterdam

Context

- Current debates on AI are just very hot!

Robots should be slaves

Joanna J. Bryson

Robots should not be described as persons, nor given legal nor moral responsibility for their actions. Robots are fully owned by us. We determine their goals and behavior, either directly or indirectly through specifying their intelligence or how their intelligence is acquired. In humanising them, we not only further dehumanise real people, but also encourage poor human decision making in the allocation of resources and responsibility. This is true at both the individual and the institutional level. This chapter describes both causes and consequences of these errors, including consequences already present in society. I make specific proposals for best incorporating robots into our society. The potential of robotics should be understood as the potential to extend our own abilities and to

Article | OPEN ACCESS

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Authors: Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell | Authors Info & Claims

FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency • Pages 610 - 623
https://doi.org/10.1145/3442188.3445922

Published: 01 March 2021 Publication History | Check for updates

99 636 366,957

Abstract

The past 3 years of work in NLP have been characterized by the development and deployment of models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, bour... sheer size. U...

In the wake of her claim, Sam Altman's "Her" tweet from May 13 has now gone viral.

Sam Altman
@sama · Follow
her
7:45 PM · May 13, 2024
40.7K Reply Copy link
Read 3K replies

Scarlett Johansson's legal team has sent two letters to OpenAI demanding clarification

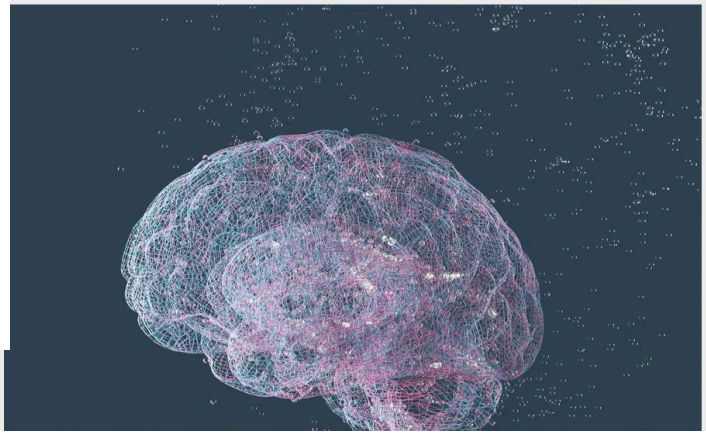
Earlier this month, OpenAI launched its latest AI personal assistant 'Sky'; and a live demonstration of its voice was held last week. Following this, many pointed out that the voice of 'Sky' sounded like Scarlett Johansson's in the 2013 romantic sci-fi film *Her*.

JULY 12, 2022 | 6 MIN READ

Google Engineer Claims AI Chatbot Is Sentient: Why That Matters

Is it possible for an artificial intelligence to be sentient?

BY LEONARDO DE COSMO



Main positions

Main positions

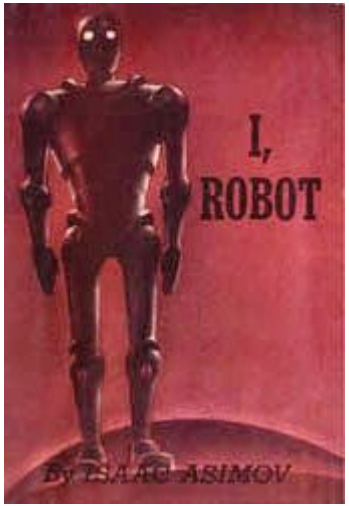
machines can NEVER be like humans

merely symbolic processors

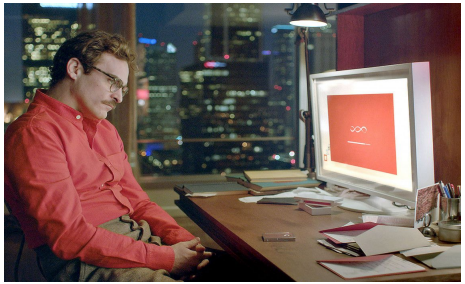
Main positions

machines can NEVER be like humans

machines can be like humans



merely symbolic processors



responsible, (possibly) sentient, (even less possibly) conscious entities

Main positions

machines are just like hammers

essentially mechanical entities

Main positions

machines are just like hammers

machines can become terminators

essentially mechanical entities

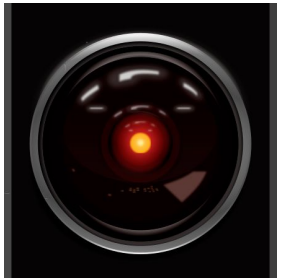
weapons with super-human capabilities

Main positions

machines are just like hammers

essentially mechanical entities

machines can become terminators*

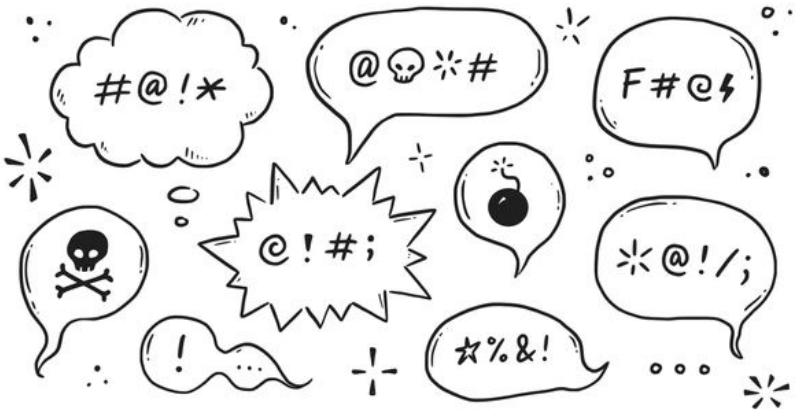


destroyers of mankind or of the world as we know it

Context

- Philosophers, technologists, legal experts, ethicists, natural scientists, engineers, CEOs, advisors, journalists, politicians… all take positions!

machines can NEVER be like humans



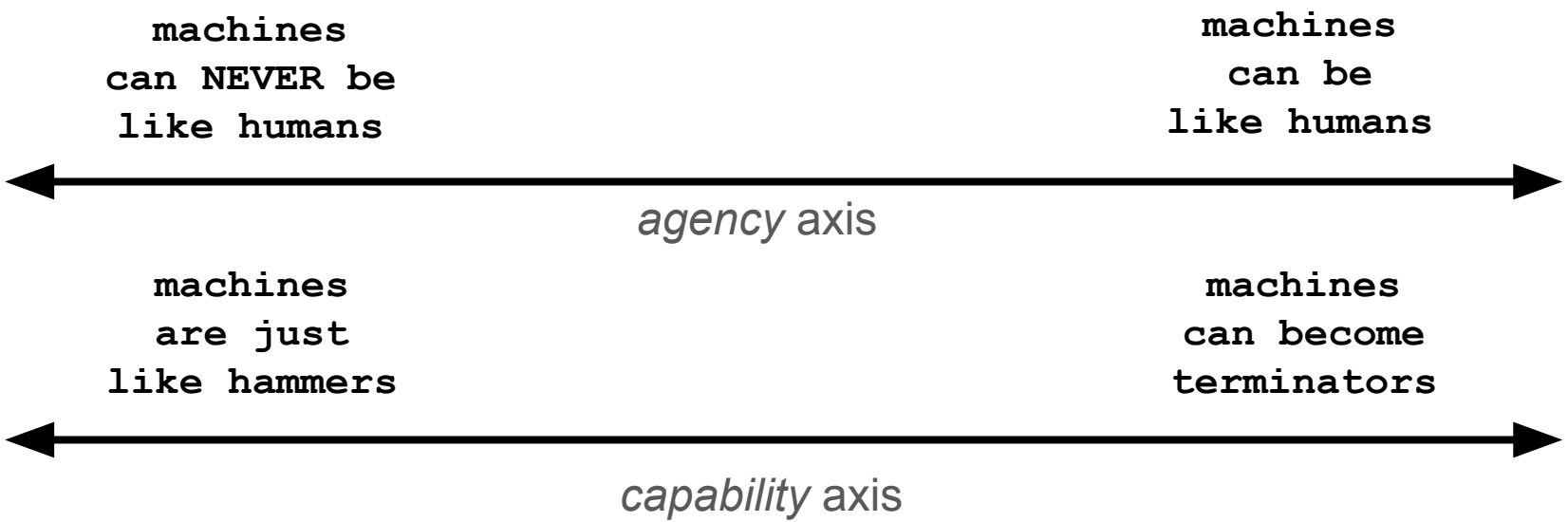
machines can be like humans

machines are just like hammers

machines can become terminators

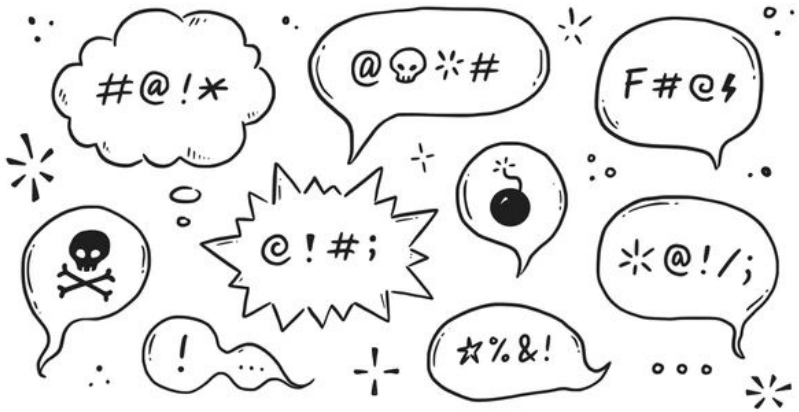
Aim of the talk

- Let us set up a framework to clarify the concepts at stake, trying to reduce the ambiguity and to unveil assumptions usually left implicit.



Starting knot!

- Consider the word “responsibility”: *moral responsibility*, *legal responsibility*, *political responsibility*, *causal responsibility*, *functional responsibility*, *(common-sense) responsibility,* and so on!

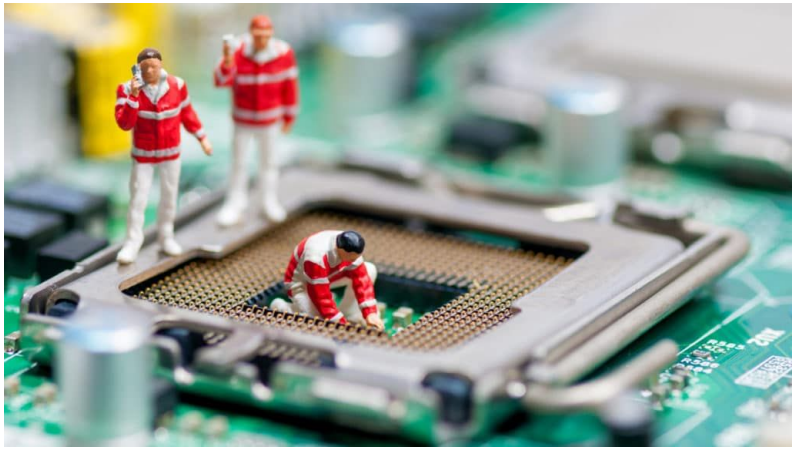


Responsibility?

- For humans, responsibility attribution is a spontaneous and seemingly universal behaviour.

Responsibility?

- For humans, responsibility attribution is a spontaneous and seemingly universal behaviour.



FUNCTION OF RESPONSIBILITY

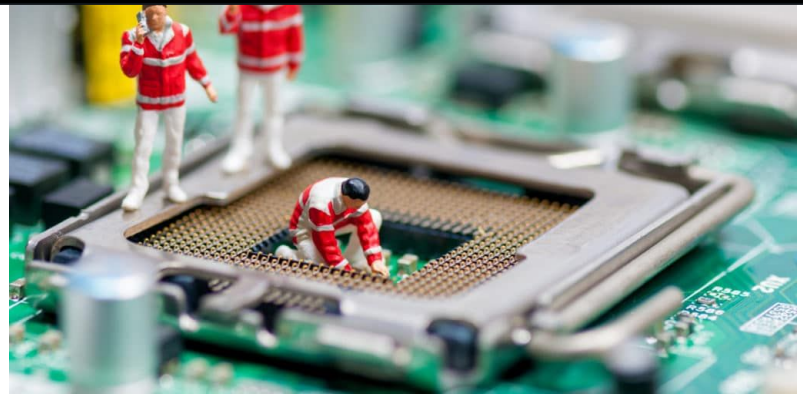
Localization of failures in wholes whose components are deemed to be independent/autonomous.

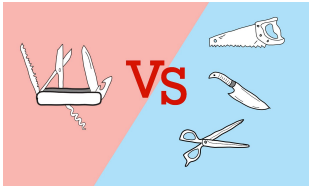
Responsibility?

Responsibility used for *computational actors* and for humans (as *moral agents*)

FUNCTION OF RESPONSIBILITY

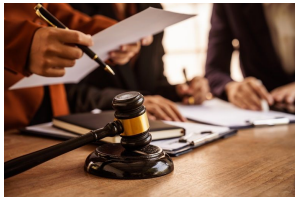
Localization of failures in wholes whose components are deemed to be independent/autonomous.





Single Responsibility Principle
in software engineering

...



Legal
responsibility

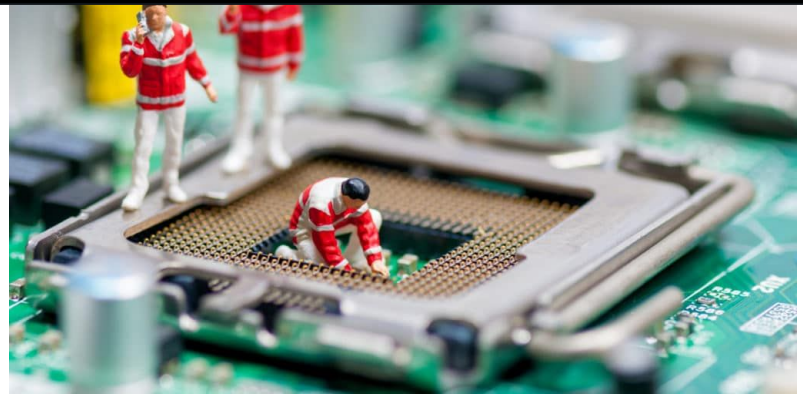
Responsibility?

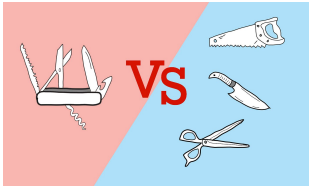
common ground: **actions!**

Responsibility used for *computational actors* and for humans (as *moral agents*)

FUNCTION OF RESPONSIBILITY

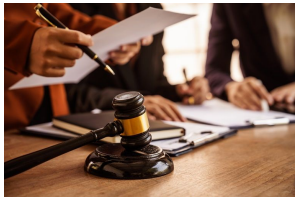
Localization of failures in wholes whose components are deemed to be independent/autonomous.





Single Responsibility Principle
in software engineering

...



Legal
responsibility

Action?

Action?

conceptualized

- It is known the same action can be described at different abstraction levels:

Brutus stabbed Caesar.
Brutus killed Caesar.
Brutus murdered Caesar.

Sowa, John F. *Knowledge representation: logical, philosophical and computational foundations*. Brooks/Cole Publishing (1999)

Action?

conceptualized

- It is known the same action can be described at different abstraction levels:

shaking hands
concluding a peace treaty
ending the war

Zacks, Jeffrey M., and Barbara Tversky. Event structure in perception and conception. Psychological bulletin 127.1 (2001): 3.

From levels of abstraction of action…

● behaviour	**how**	shaking hands
● outcome	**what**	concluding a peace treaty
● policy	**why**	ending the war

…to levels of responsibility!

● behaviour	**how**	operational responsibility
● outcome	**what**	tactical responsibility
● policy	**why**	strategic responsibility

…to levels of responsibility!

- behaviour **how** — operational responsibility
- outcome **what** — tactical responsibility
- policy **why** — strategic responsibility

a component may fail in each of these:

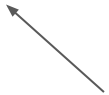
- behaviour: not performing what it is expected to
- outcome: not achieving what it is expected to
- policy: not abiding by what it is expected to, *while achieving the what*

…to levels of responsibility!

- behaviour **how** — operational responsibility
- outcome **what** — tactical responsibility
- policy **why** — strategic responsibility

a component may fail in each of these:

- behaviour: not performing what it is expected to
- outcome: not achieving what it is expected to
- policy: not abiding by what it is expected to, *while achieving the what*



this is at a second-level!

Example 1

- **goal**: fishing
- **reward**: proportional to quantity of fish, inversely to effort.

solution to optimization problem

Example 1

- **goal**: fishing
- **reward**: proportional to quantity of fish, inversely to effort.

solution to optimization problem

fishing with bombs

Example 1

- **goal**: fishing
- **reward**: proportional to quantity of fish, inversely to effort.

solution to optimization problem

fishing with bombs

no problem with **behaviour**

Example 1

- **goal**: fishing
- **reward**: proportional to quantity of fish, inversely to effort.

solution to optimization problem

no problem with **outcome**

fishing with bombs

no problem with **behaviour**

Example 1

- **goal**: fishing
- **reward**: proportional to quantity of fish, inversely to effort.

no problem with **outcome**

solution to optimization problem

fishing with bombs

serious problems with **policy**!

no problem with **behaviour**

Example 2

You are asked to help the police to identify venues of synthetic drug production.

- Synthetic drug is usually produced in barns rented for a few months, then abandoned, and chemical residuals thrown in the canals.
- Agriculture is not rentable at the moment, barn owners may be more lenient in checking who is renting their barn

Example 2

You are asked to help the police to identify venues of synthetic drug production.

- Synthetic drug is usually produced in barns rented for a few months, then abandoned, and chemical residuals thrown in the canals.
- Agriculture is not rentable at the moment, barn owners may be more prompt in checking who is renting their barn

*let us build a **risk indicator** for the police*: **if an area is becoming poorer we may expect barns be rented for drug production**

Example 2

You are asked to help the police to identify venues of synthetic drug production.

- Synthetic drug is usually produced in barns rented for a few months, then abandoned, and chemical residuals thrown in the canals.
- Agriculture is not rentable at the moment, barn owners may be more indifferent in checking who is renting their barn

*let us build a **risk indicator** for the police*: **if an area is becoming poorer we may expect barns be rented for drug production**

Example 2

You are asked to help the police to identify venues of synthetic drug production.

- Synthetic drug is usually produced in barns rented for a few months, then abandoned, and chemical residuals thrown in the canals.

Both examples show the difficulties of aligning policies with outcomes!

→ *let us build a **risk indicator** for the police*: **if an area is becoming poorer we may expect barns be rented for drug production**

Can machines
be like humans?

Domains of responsibility

- **responsibility** to the agent(s) determining the action to occur

Domains of responsibility

- **responsibility** to the agent(s) determining the action to occur
- **accountability** to the agent(s) justifying the action to occur
  - *ex-ante* or process-level: *auditability*, *compliance checking*
  - *ex-post* or event-level: *forensics, judiciary activity*

Domains of responsibility

- **responsibility** to the agent(s) determining the action to occur
- **accountability** to the agent(s) justifying the action to occur
  - *ex-ante* or process-level: *auditability*, *compliance checking*
  - *ex-post* or event-level: *forensics, judiciary activity*
- **liability** to the agent(s) be blamed or praised for the action

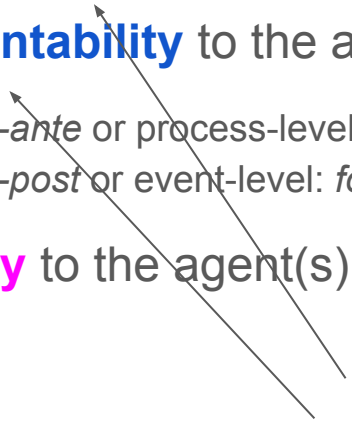
Domains of responsibility

- **responsibility** to the agent(s) determining the action to occur
- **accountability** to the agent(s) justifying the action to occur
  - *ex-ante* or process-level: *auditability*, *compliance checking*
  - *ex-post* or event-level: *forensics, judiciary activity*
- **liability** to the agent(s) be blamed or praised for the action

e.g. in law, accountability covers the three domains:
act responsibly** (behave following the rules) and **take responsibility

Domains of responsibility

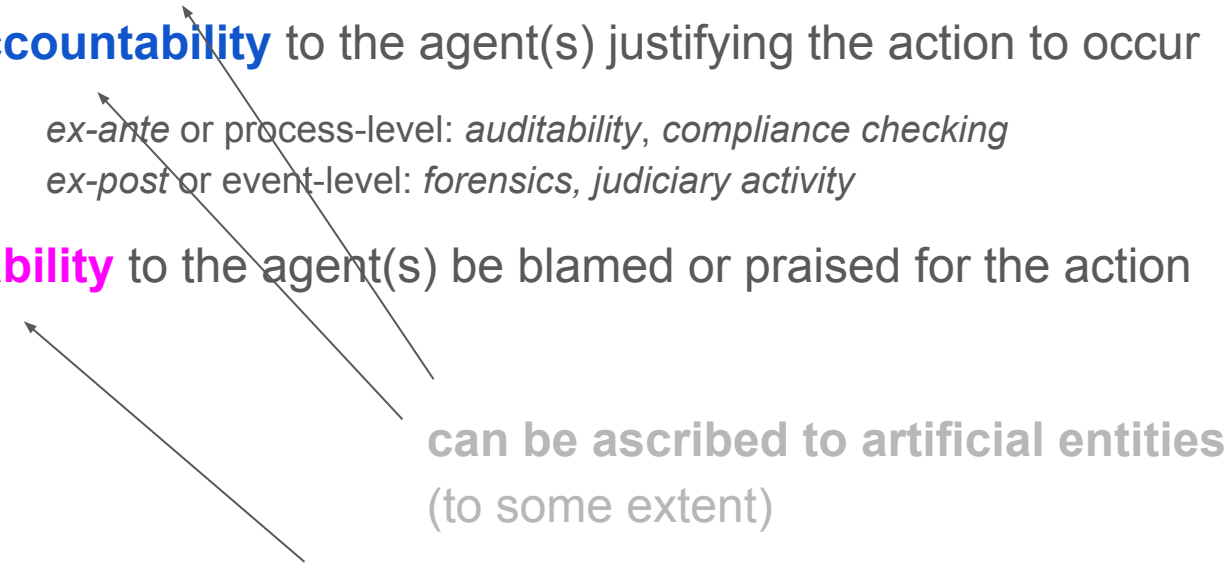
- **responsibility** to the agent(s) determining the action to occur
- **accountability** to the agent(s) justifying the action to occur
  - *ex-ante* or process-level: *auditability*, *compliance checking*
  - *ex-post* or event-level: *forensics, judiciary activity*
- **liability** to the agent(s) be blamed or praised for the action



can be ascribed to artificial entities
(to some extent)

Domains of responsibility

- **responsibility** to the agent(s) determining the action to occur
- **accountability** to the agent(s) justifying the action to occur
  - *ex-ante* or process-level: *auditability*, *compliance checking*
  - *ex-post* or event-level: *forensics, judiciary activity*
- **liability** to the agent(s) be blamed or praised for the action



can be ascribed to artificial entities
(to some extent)

is always upon humans

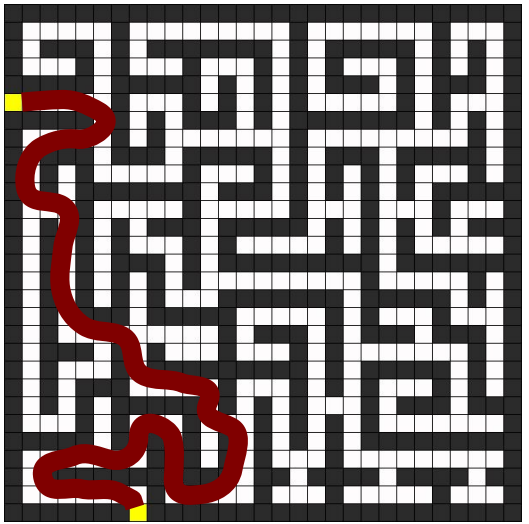
Levels of responsibility

- behaviour **how** operational responsibility

Levels of responsibility

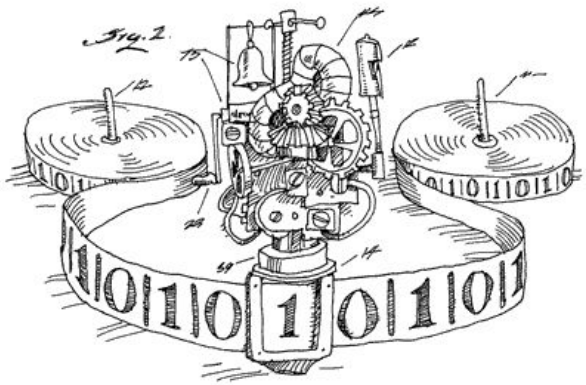
- behaviour **how** operational responsibility

I expect you to apply specific instructions



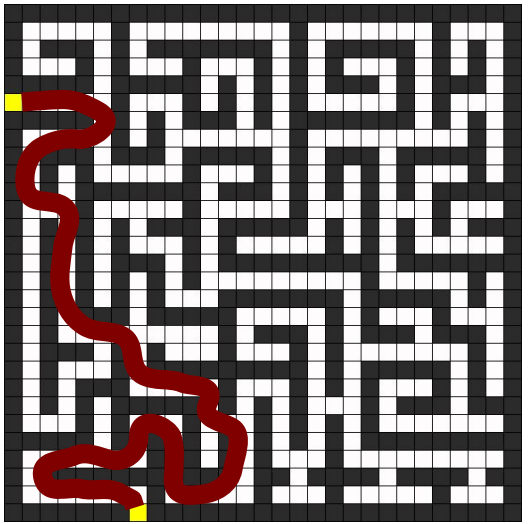
Levels of responsibility

- behaviour **how** operational responsibility



traditional domain of **informatics,** and engineering at large

I expect you to apply specific instructions



Levels of responsibility

- behaviour **how** operational responsibility
- outcome **what** tactical responsibility

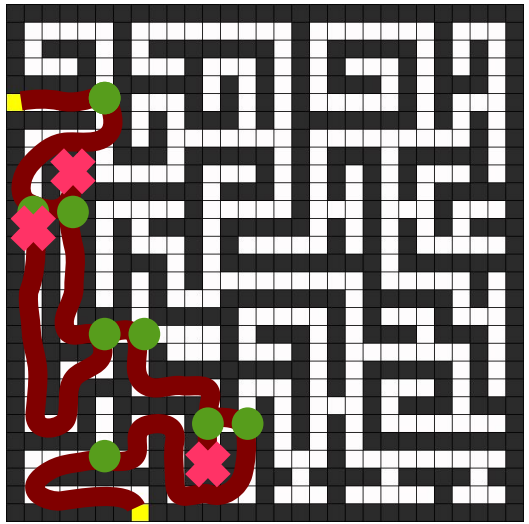
Levels of responsibility

- behaviour **how**
- outcome **what**

operational responsibility

tactical responsibility

I expect you to
be successful
in a specific task



Levels of responsibility

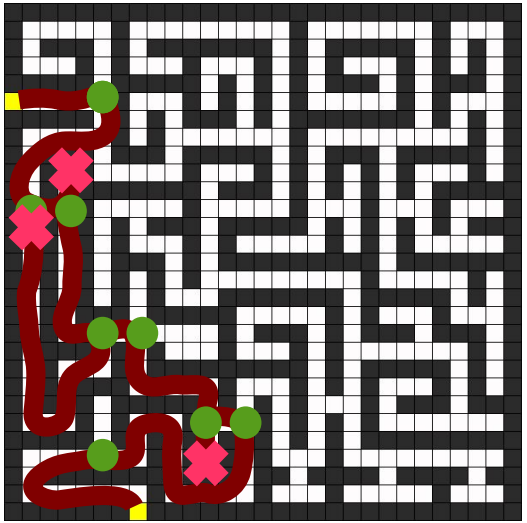
- behaviour **how**
- outcome **what**

operational responsibility

tactical responsibility

traditional domain of **AI**

I expect you to be successful in a specific task

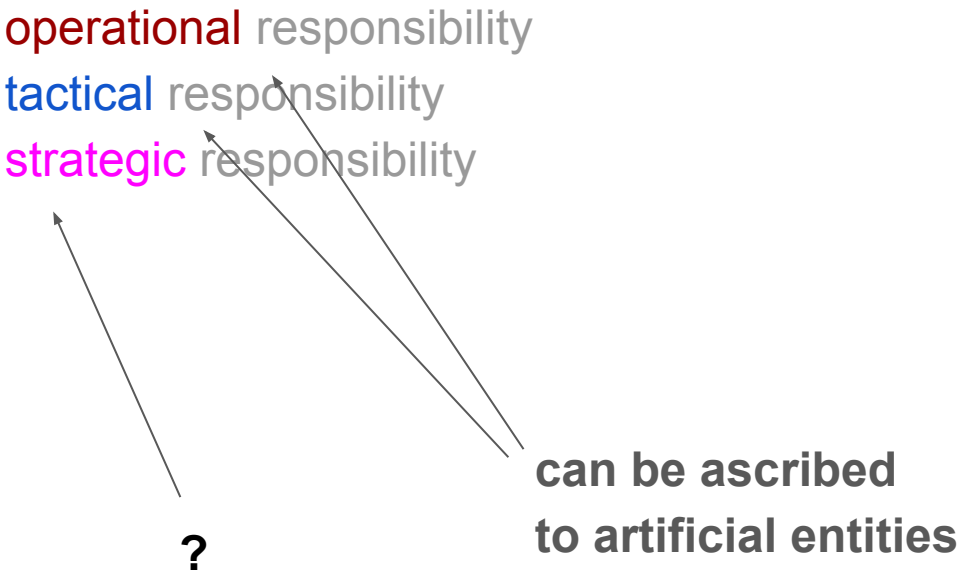


Levels of responsibility

● behaviour	**how**	operational responsibility
● outcome	**what**	tactical responsibility
● policy	**why**	strategic responsibility

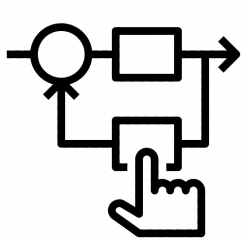
Levels of responsibility

- behaviour **how**
- outcome **what**
- policy **why**



Requirements for strategic responsibility

An agent has *strategic responsibility* if it:



has the ability to **control** its own behaviour

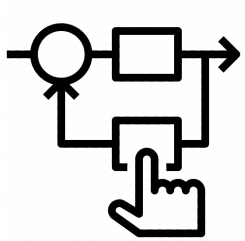
has the ability to **foresee** the associated outcomes

has the ability to **evaluate** actions according to a certain normative/value structure

Sileno, G., Boer, A., Gordon, G., Rieder, B., Like Circles in the Water: Responsibility as a System-Level Function. Proceedings of 3rd XAILA workshop: Explainable and Responsible AI and Law, in conjunction with JURIX 2020 (2020)

Requirements for strategic responsibility

An agent has *strategic responsibility* if it:



has the ability to **control** its own behaviour

has the ability to **foresee** the associated outcomes

necessary e.g. to identify wrong behaviour

has the ability to **evaluate** actions according to a certain normative/value structure

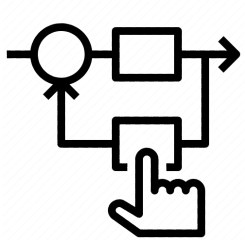
Sileno, G., Boer, A., Gordon, G., Rieder, B., Like Circles in the Water: Responsibility as a System-Level Function. Proceedings of 3rd XAILA workshop: Explainable and Responsible AI and Law, in conjunction with JURIX 2020 (2020)

Requirements for strategic responsibility

An agent has *strategic responsibility* if it:

necessary e.g. to inhibit wrong behaviour

necessary e.g. to identify wrong behaviour



has the ability to **control** its own behaviour

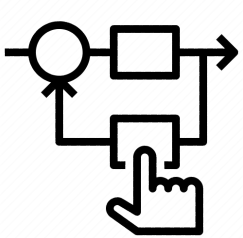
has the ability to **foresee** the associated outcomes

has the ability to **evaluate** actions according to a certain normative/value structure

Sileno, G., Boer, A., Gordon, G., Rieder, B., Like Circles in the Water: Responsibility as a System-Level Function. Proceedings of 3rd XAILA workshop: Explainable and Responsible AI and Law, in conjunction with JURIX 2020 (2020)

Requirements for strategic responsibility

An agent has *strategic responsibility* if it:



has the ability to **control** its own behaviour

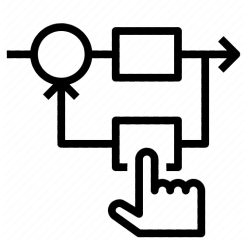
has the ability to **foresee** the associated outcomes

has the ability to **evaluate** actions according to a certain normative/value structure

can be ascribed to artificial entities
(to some extent)

Requirements for strategic responsibility

An agent has *strategic responsibility* if it:



has the ability to **control** its own behaviour

has the ability to **foresee** the associated outcomes

?

has the ability to **evaluate** actions according to a certain normative/value structure

can be ascribed to artificial entities
(to some extent)

Evaluative ability

The process of evaluation can be decomposed into:

Santoni de Sio F, van den Hoven J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI. 2018;5.

Evaluative ability

The process of evaluation can be decomposed into:

- **basic level** (***applied morality***), specifying:
  - *content*, ie. situations and actions to be evaluated
  - *criteria*, ie. the basis against which to perform the evaluation
  - *process*: ie. how (heuristics, or procedure) to evaluate
  - *acceptance conditions*, ie. when a particular behaviour is acceptable

Santoni de Sio F, van den Hoven J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI. 2018;5.

Evaluative ability

The process of evaluation can be decomposed into:

- **basic level** (***applied morality***), specifying:
  - *content*, ie. situations and actions to be evaluated
  - *criteria*, ie. the basis against which to perform the evaluation
  - *process*: ie. how (heuristics, or procedure) to evaluate
  - *acceptance conditions*, ie. when a particular behaviour is acceptable
- **meta-level** (***volitional morality***): how, and on the basis of what we define components at the basic level.

Santoni de Sio F, van den Hoven J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI. 2018;5.

Evaluative ability

can be ascribed to artificial entities (to some extent)

The process of evaluation can be decomposed into:

- **basic level** (***applied morality***), specifying:
  - *content*, ie. situations and actions to be evaluated
  - *criteria*, ie. the basis against which to perform the evaluation
  - *process*: ie. how (heuristics, or procedure) to evaluate
  - *acceptance conditions*, ie. when a particular behaviour is acceptable
- **meta-level** (***volitional morality***): how, and on the basis of what we define components at the basic level.

Santoni de Sio F, van den Hoven J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI. 2018;5.

Evaluative ability

can be ascribed to artificial entities (to some extent)

The process of evaluation can be decomposed into:

- **basic level** (***applied morality***), specifying:
  - *content*, ie. situations and actions to be evaluated
  - *criteria*, ie. the basis against which to perform the evaluation
  - *process*: ie. how (heuristics, or procedure) to evaluate
  - *acceptance conditions*, ie. when a particular behaviour is acceptable
- **meta-level** (***volitional morality***): how, and on the basis of what we define components at the basic level.

domain of **ethics** (for morality) and **jurisprudence** (for legality) **eventually lies upon humans**

Evaluative ability

can be ascribed to artificial entities (to some extent)

The process of evaluation can be decomposed into:

- **basic level** (***applied morality***), specifying:
  - *content*, ie. situations and actions to be evaluated

The highest hierarchical levels of the evaluative framework for strategic responsibility are always human matter

- **meta-level** (***volitional morality***): how, and on the basis of what we define components at the basic level.

domain of **ethics** (for morality) and **jurisprudence** (for legality) **eventually lies upon humans**

Can machines
be terminators?

Can machines be terminators?

- An autonomous weapon may have better control and better foreseeability than humans, yet its evaluation components today provide very limited behavioural boundaries…

Can machines be terminators?

- An autonomous weapon may have better control and better foreseeability than humans, yet its evaluation components today provide very limited behavioural boundaries…

→ it is a problematic device!

Can machines be terminators?

- An autonomous weapon may have better control and better foreseeability than humans, yet its evaluation components today provide very limited behavioural boundaries…

→ it is a problematic device!

- This is even more the case for the ***paperclip maximizer***: a single concrete objective, which may be realized in a coalition of artificial entities, with effective control.

Can machines be terminators?

LOCALITY OF ACTION

- An autonomous weapon may have better control and better foreseeability than humans, yet its evaluation components today provide very limited behavioural boundaries…

→ it is a problematic device!

- This is even more the case for the ***paperclip maximizer***: a single concrete objective, which may be realized in a coalition of artificial entities, with effective control.

NON-LOCALITY OF ACTION

But wait…

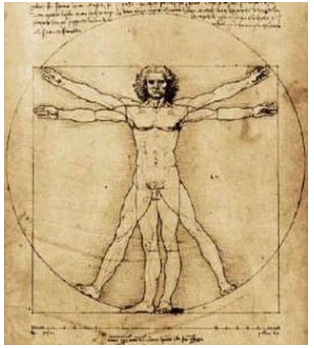
- General intelligence is NOT the core issue here!

NON-LOCALITY OF ACTION

But wait…

- General intelligence is NOT the core issue here!
- Indeed, we humans have general intelligence,
  yet we are not terminators…

  aren't we?

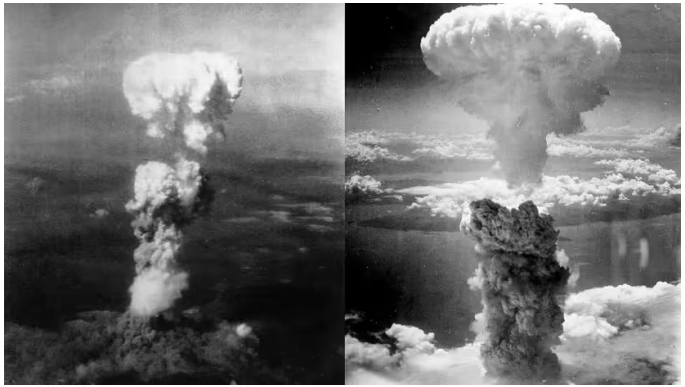


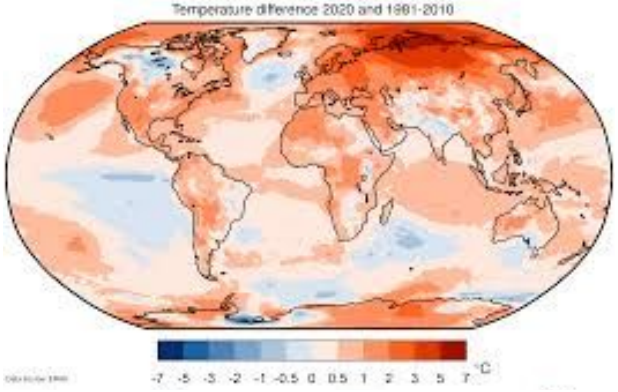
LOCALITY OF ACTION

NON-LOCALITY OF ACTION

Can humans be terminators?

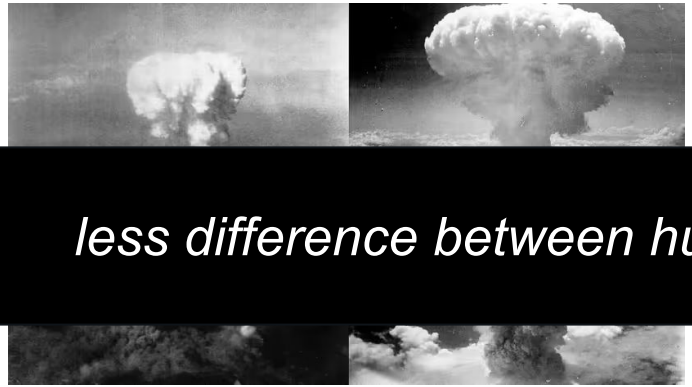
- We humans approach the terminator role when our intervention become much more impactful than what we were evolutionary selected to be:
  - at individual level, eg. atomic bombs
  - at collective level, eg. pollution, and then climate warming

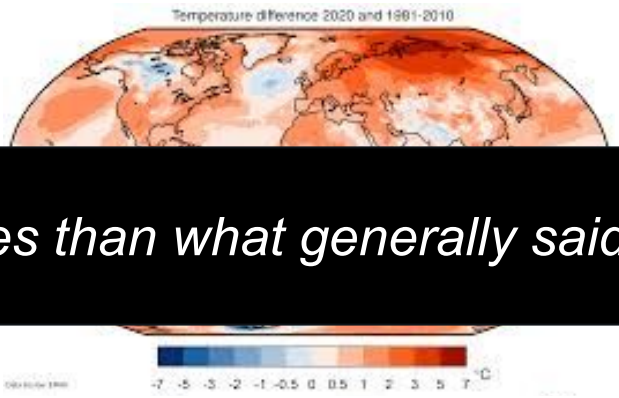




Can humans be terminators?

- We humans approach the terminator role when our intervention become much more impactful than what we were evolutionary selected to be:
  - at individual level, eg. atomic bombs
  - at collective level, eg. pollution, and then climate warming





less difference between humans and machines than what generally said

General responsibility principle

- The more the entity has control,

  (ie. it is able to perform impactful actions),

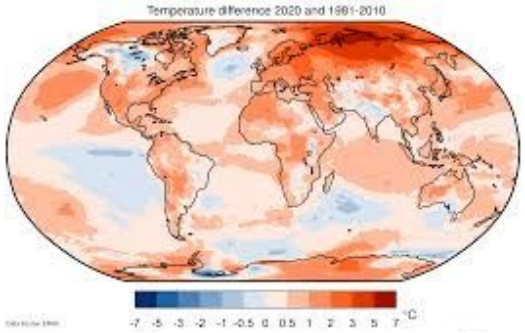
- The more it requires foreseeability,

  (ie. it is able to predict the impact it may produce)

- The more it requires an adequate evaluation structure

  (eg. socially acceptable and sustainable).

Humans and climate warming



- The more the entity has control,

  (ie. it is able to perform impactful actions),

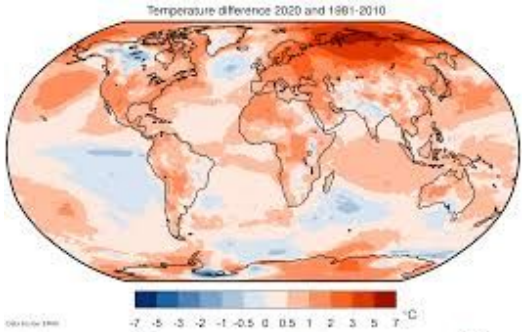
- The more it requires foreseeability, and

  (ie. it is able to predict the impact it may produce)

- The more it requires an adequate evaluation structure

  (eg. socially acceptable and sustainable).

Humans and climate warming



- The more the entity has control,

  (ie. it is able to perform impactful actions),

- The more it requires foreseeability, and

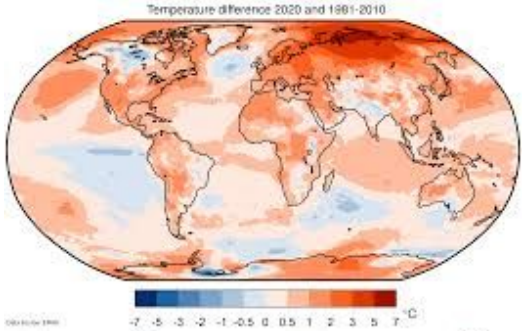
  (ie. it is able to predict the impact it may produce)

- The more it requires an adequate evaluation structure

  (eg. socially acceptable and sustainable).

We started having impact decades go.

Humans and climate warming



- The more the entity has control,

  (ie. it is able to perform impactful actions),

- The more it requires foreseeability, and

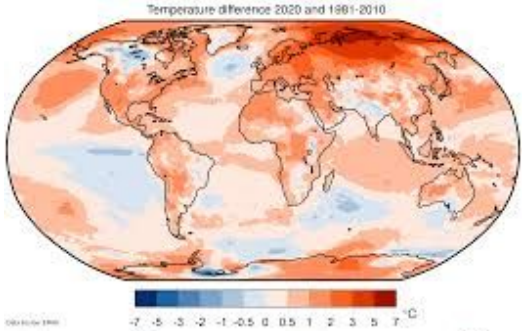
  (ie. it is able to predict the impact it may produce)

- The more it requires an adequate evaluation structure

  (eg. socially acceptable and sustainable).

We started having impact decades go. We are more and more becoming aware of what will happen.

Humans and climate warming



- The more the entity has control,

  (ie. it is able to perform impactful actions),

- The more it requires foreseeability, and

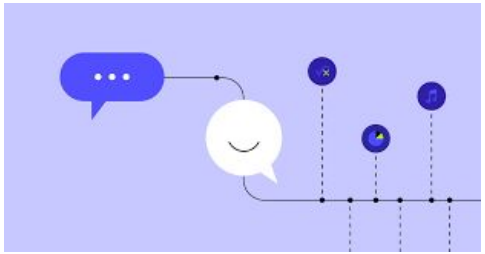
  (ie. it is able to predict the impact it may produce)

- The more it requires an adequate evaluation structure

  (eg. socially acceptable and sustainable).

We started having impact decades go. We are more and more becoming aware of what will happen. At level of policy, still very slow.

Online chatbots based on LLMs



- The more the entity has control,

  (ie. it is able to perform impactful actions),

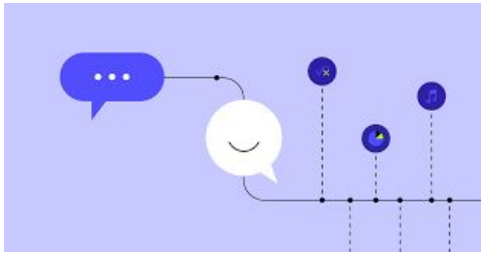
- The more it requires foreseeability, and

  (ie. it is able to predict the impact it may produce)

- The more it requires an adequate evaluation structure

  (eg. socially acceptable and sustainable).

Online chatbots based on LLMs



- The more the entity has control,

  (ie. it is able to perform impactful actions),

- The more it requires foreseeability, and

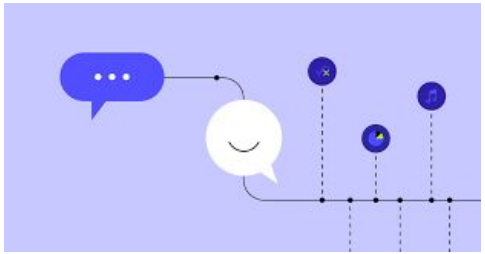
  (ie. it is able to predict the impact it may produce)

- The more it requires an adequate evaluation structure

  (eg. socially acceptable and sustainable).

→ Online chatbots interact globally.

Online chatbots based on LLMs



- The more the entity has control,

  (ie. it is able to perform impactful actions),

- The more it requires foreseeability, and

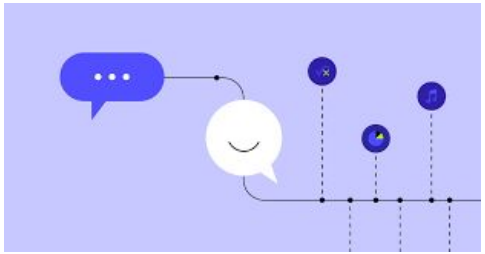
  (ie. it is able to predict the impact it may produce)

- The more it requires an adequate evaluation structure

  (eg. socially acceptable and sustainable).

→ Online chatbots interact globally. They are trained against this continuous feed and other unknown inputs.

Online chatbots based on LLMs



- The more the entity has control,

  (ie. it is able to perform impactful actions),

- The more it requires foreseeability, and

  (ie. it is able to predict the impact it may produce)

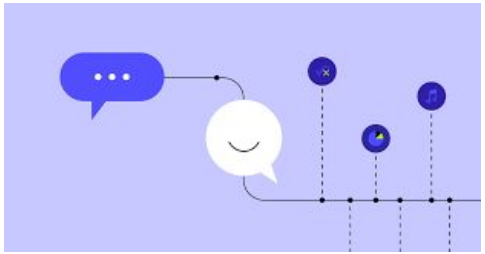
- The more it requires an adequate evaluation structure

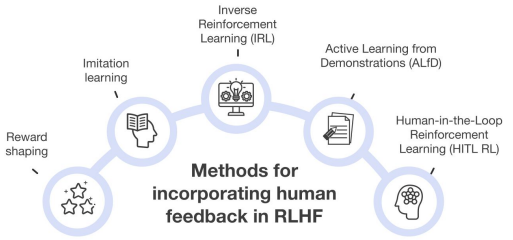
  (eg. socially acceptable and sustainable).

→ Online chatbots interact globally. They are trained against this continuous feed and other unknown inputs. What about their policy???

Online chatbots based on LLMs: policy level

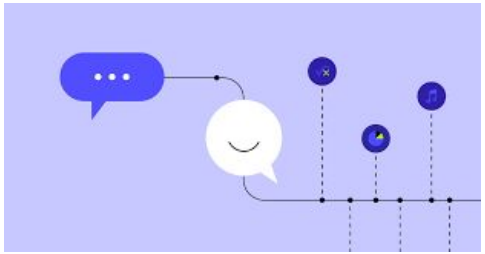
- From a technical point of view, chatbots are fine-tuned via *Reinforcement Learning from Human Feedback* (**RLHF**), to e.g. minimize harmful or untruthful outputs.

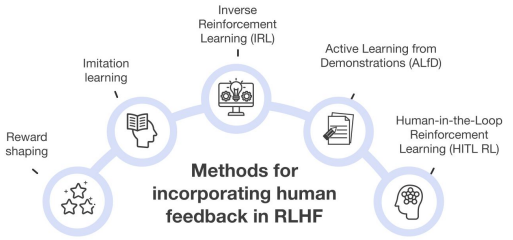




Online chatbots based on LLMs: policy level

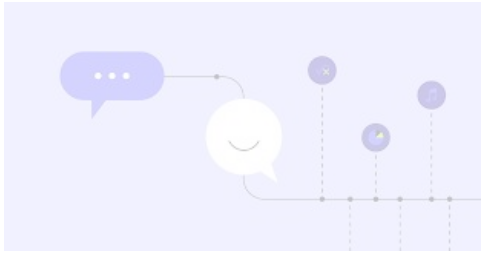
- From a technical point of view, chatbots are fine-tuned via *Reinforcement Learning from Human Feedback* (**RLHF**), to e.g. minimize harmful or untruthful outputs.
- Arguable whether ***mimicking human preferences*** is the best way to achieve moral behaviour.





Online chatbots based on LLMs: policy level

- From a technical point of view, chatbots are fine-tuned via *Reinforcement Learning from Human Feedback* (**RLHF**), to e.g. minimize harmful or untruthful outputs.
- Arguable whether ***mimicking human preferences*** is the best way to achieve moral behaviour.

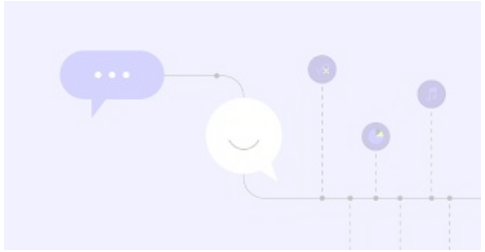


vs Human social systems: policy level

- Humans define policies using **top-down** processes (concerning mostly ***legality***) with **bottom-up** processes (concerning mostly ***legitimacy***)

Online chatbots based on LLMs: policy level

- From a technical point of view, chatbots are fine-tuned via *Reinforcement Learning from Human Feedback* (**RLHF**), to e.g. minimize harmful or untruthful outputs.



Current chatbots miss the top-down, and fail locality for the bottom-up

vs Human social systems: policy level

- Humans define policies using **top-down** processes (concerning mostly ***legality***) with **bottom-up** processes (concerning mostly ***legitimacy***)

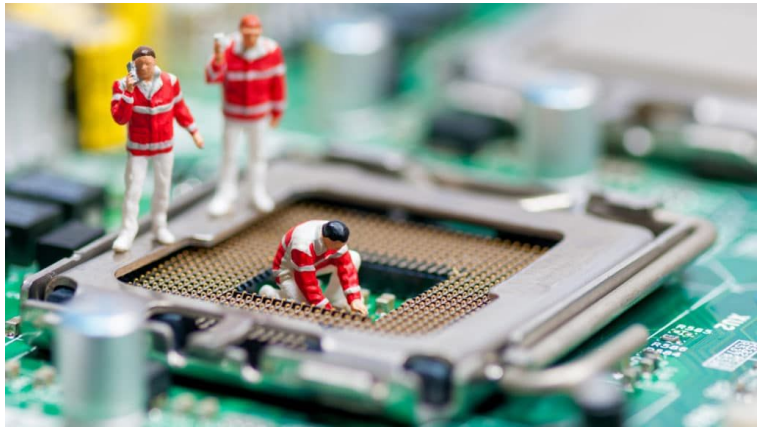
Some conclusions

Conclusions (1)

- Humans are always **eventually responsible** (at least on a policy level),
- Humans are the **only ones that can be liable.**

Conclusions (1)

- Humans are always eventually responsible (at least on a policy level),
- Humans are the only ones that can be liable.
- Machines can only — and when used, they should — cover **lower levels of responsibility and accountability.**



Conclusions (2)

- To any increase of **control**, and **foreseeability** skills, there needs to be adequate modifications on the **evaluative framework** side.

Conclusions (2)

- To any increase of **control,** and **foreseeability** skills, there needs to be adequate modifications on the **evaluative framework** side.
- For artificial systems, that means we need to think in terms of **computational policy mechanisms**

Conclusions (2)

- To any increase of **control**, and **foreseeability** skills, there needs to be adequate modifications on the **evaluative framework** side.
- For artificial systems, that means we need to think in terms of **computational policy mechanisms** ⇒ eg. the call for **NORMWARE**

HARDWARE

SOFTWARE

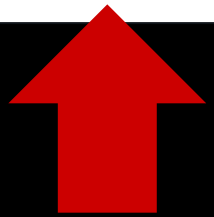
?

NORMWARE

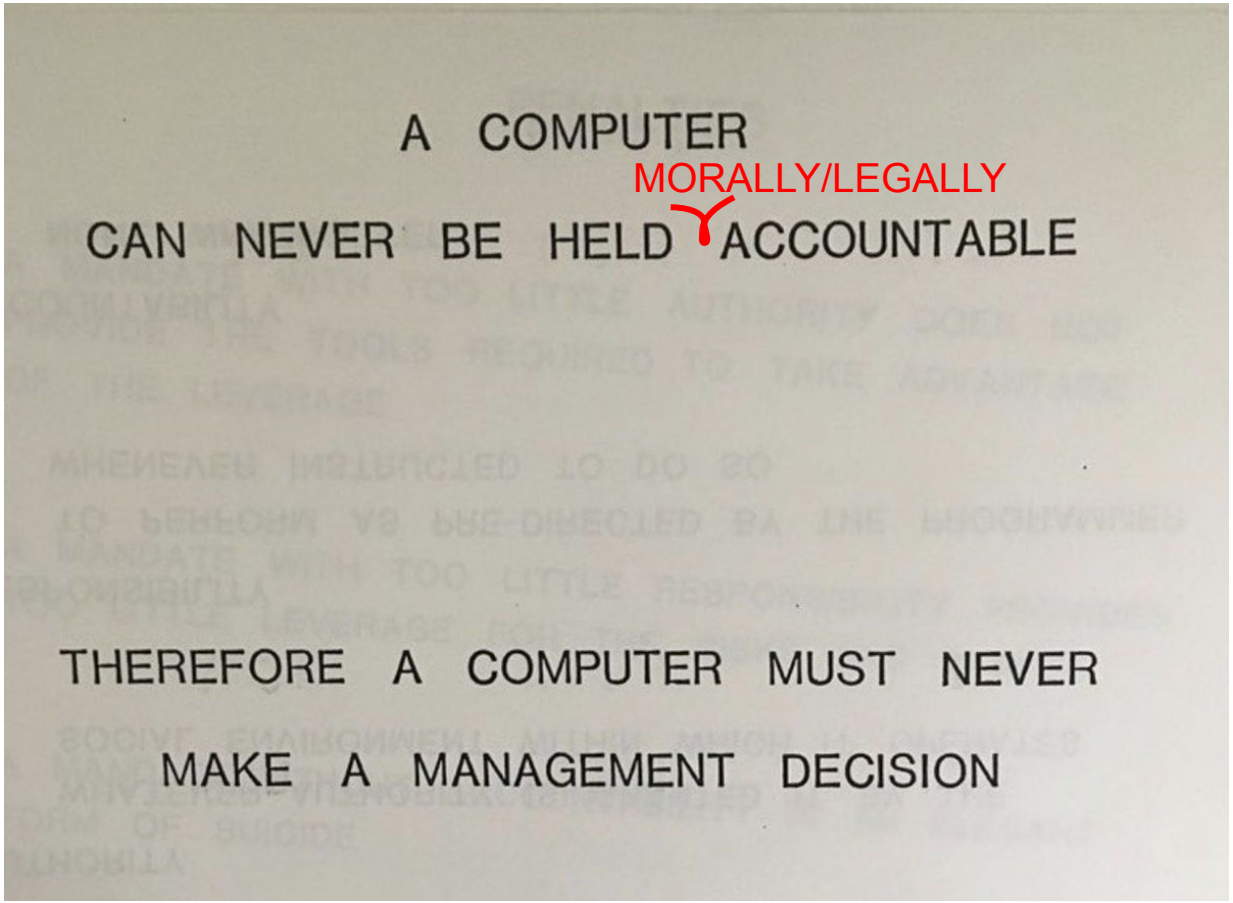
Sileno, G., Boer, A. and van Engers, T., The Role of Normware in Trustworthy and Explainable AI, Proceedings of XAILA workshop: Explainable AI and Law, in conjunction with JURIX 2018.
Sileno, G., *Code-driven law NO, Normware SI!,* presented at Conference on Cross-disciplinary Research in Computational Law (CRCL 2022), 2022. https://arxiv.org/pdf/2410.17257

Conclusions (2)

- To any increase of **control**, and **foreseeability** skills, there needs to be adequate modifications on the **evaluative framework** side.



If we cannot guarantee this last part,
better no increase in the first two dimensions!



IBM slide from 1979.

based on Sileno, G., Zurek, T., Between Hammer and Terminator,
Workshop on Artificial Intelligence, Law, and Philosophy, joint to JURIX 2023 (ALP2023).

Between Hammer and Terminator

BNAIC, invited talk at FACT session
Utrecht University, 19th November 2024

Giovanni Sileno
g.sileno@uva.nl
University of Amsterdam

joint work with Tomasz Zurek
t.a.zurek@uva.nl
University of Amsterdam