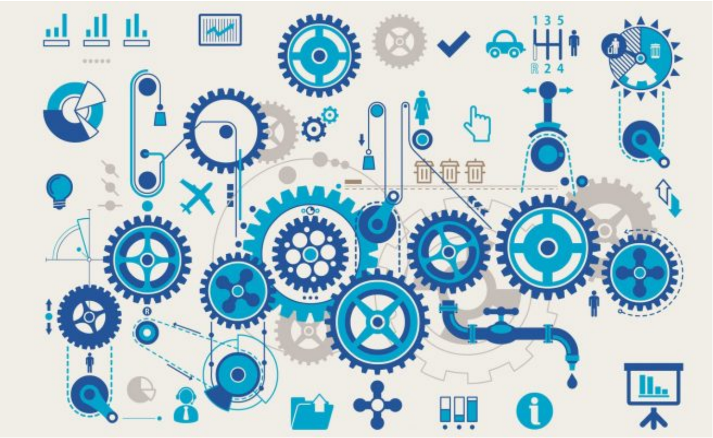
A Computational Model of Moral and Legal Responsibility via Simplicity Theory

Giovanni SILENO [a,1], Antoine SAILLENFEST [b] and Jean-Louis DESSALLES [a]
[a] *LTCI, Télécom ParisTech, Université Paris-Saclay, 46 rue Barrault, Paris, France*
[b] *Geronimo Agency, 33 rue d'Artois, Paris, France*

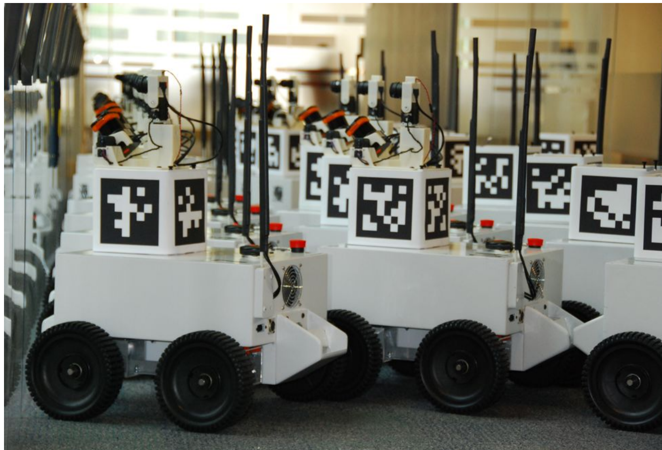
giovanni.sileno@telecom-paristech.fr

15 December 2017, JURIX @ Luxembourg



with the (supposedly) near advent of ***autonomous artificial entities***, or similar forms of ***distributed automatic decision making***,

to define ***operationally*** the notion of **responsibility** becomes of primary importance.



How to compute responsibility?

- Traditional research track in AI & Law:

How to compute responsibility?

- Traditional research track in AI & Law:
  - structural (logical) approaches
    - focus on **reasoning constructs**: Ontologies [Lehmann et al., 2004], Inferences [Prakken, 2002] or Stories [Bex et al., 2000]

How to compute responsibility?

- Traditional research track in AI & Law:
  - structural (logical) approaches
    - focus on **reasoning constructs**: Ontologies [Lehmann et al., 2004], Inferences [Prakken, 2002] or Stories [Bex et al., 2000]
  - quantitative approaches
    - focus on **relative support of evidence**: Bayesian inference [Fenton et al., 2012], Causal Bayesian Networks [Halpern, 2015]

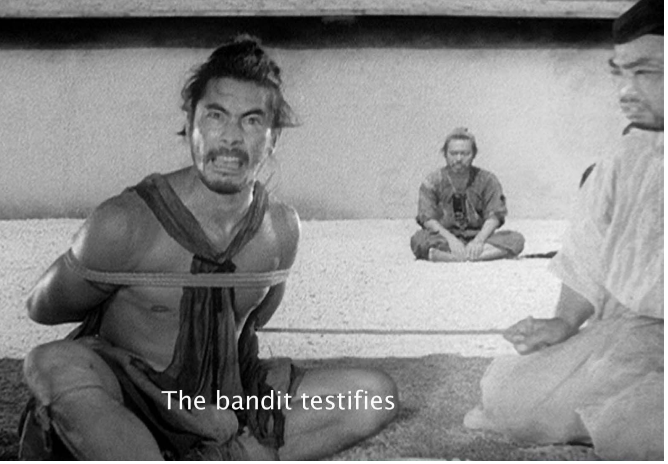
How to compute responsibility?

- Traditional research track in AI & Law:
  - structural (logical) approaches
    - focus on **reasoning constructs**: Ontologies [Lehmann et al., 2004], Inferences [Prakken, 2002] or Stories [Bex et al., 2000]
  - quantitative approaches
    - focus on **relative support of evidence**: Bayesian inference [Fenton et al., 2012], Causal Bayesian Networks [Halpern, 2015]
  - hybrid methods [Vlek et al., 2014], [Verheij, 2014]

How to compute responsibility?

- Traditional research track in AI & Law:
  - structural (logical) approaches
    - focus on **reasoning constructs**: Ontologies [Lehmann et al., 2004], Inferences [Prakken, 2002] or Stories [Bex et al., 2000]
  - quantitative approaches
    - focus on **relative support of evidence**: Bayesian inference [Fenton et al., 2012], Causal Bayesian Networks [Halpern, 2015]
  - hybrid methods [Vlek et al., 2014], [Verheij, 2014]

- Here we introduce an alternative research direction, building upon **cognitive models**.

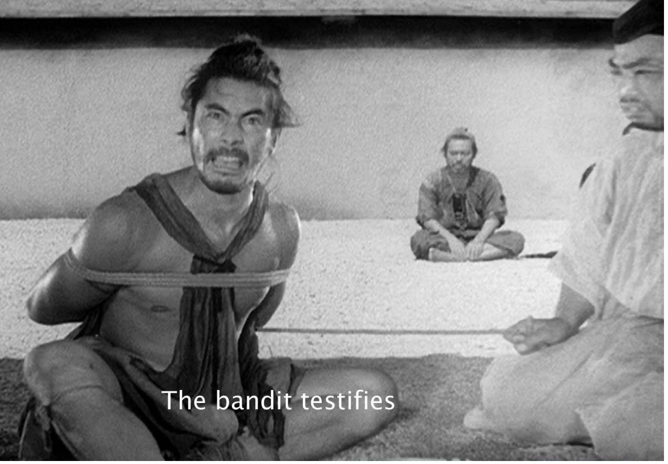



Rashomon, 1950


12 Angry Men, 1956

Responsibility attribution for humans

- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.

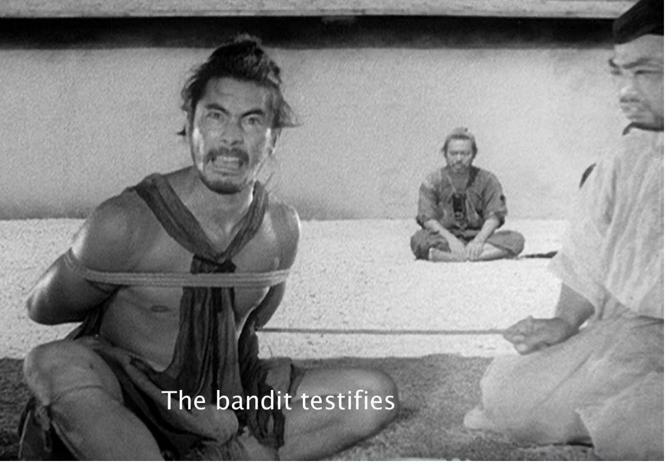


Rashomon, 1950

12 Angry Men, 1956

Responsibility attribution for humans

- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.
- Non-related ancient legal systems bear much resemblance to modern law and seem perfectly sensible nowadays.



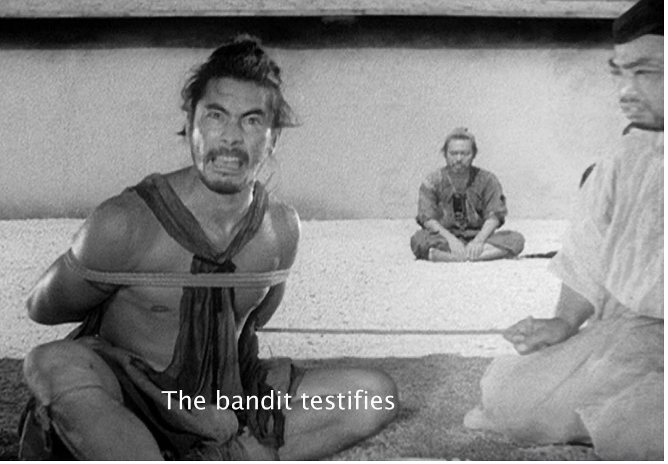
Rashomon, 1950

12 Angry Men, 1956

Responsibility attribution for humans

- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.
- Non-related ancient legal systems bear much resemblance to modern law and seem perfectly sensible nowadays.

→ *responsibility attribution* may be controlled by fundamental cognitive mechanisms.




Rashomon, 1950

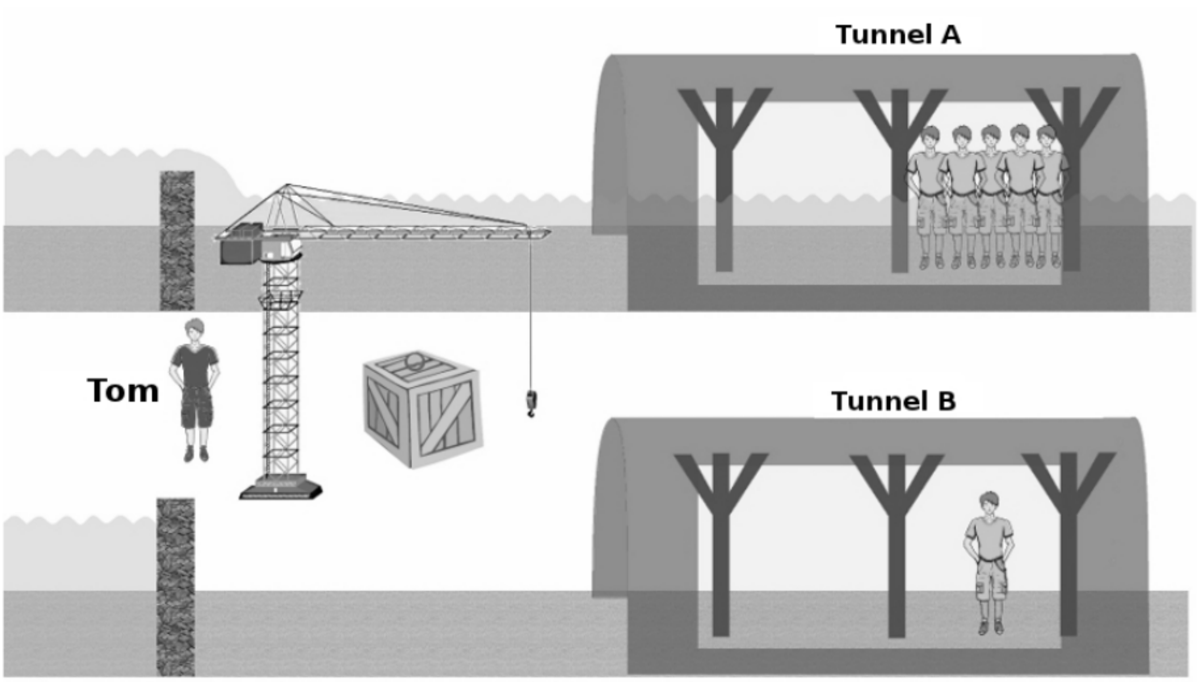
12 Angry Men, 1956

Responsibility attribution for humans

- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.
- Non-related ancient legal systems bear much resemblance to modern law and seem perfectly sensible nowadays.

→ *responsibility attribution* may be controlled by fundamental cognitive mechanisms.

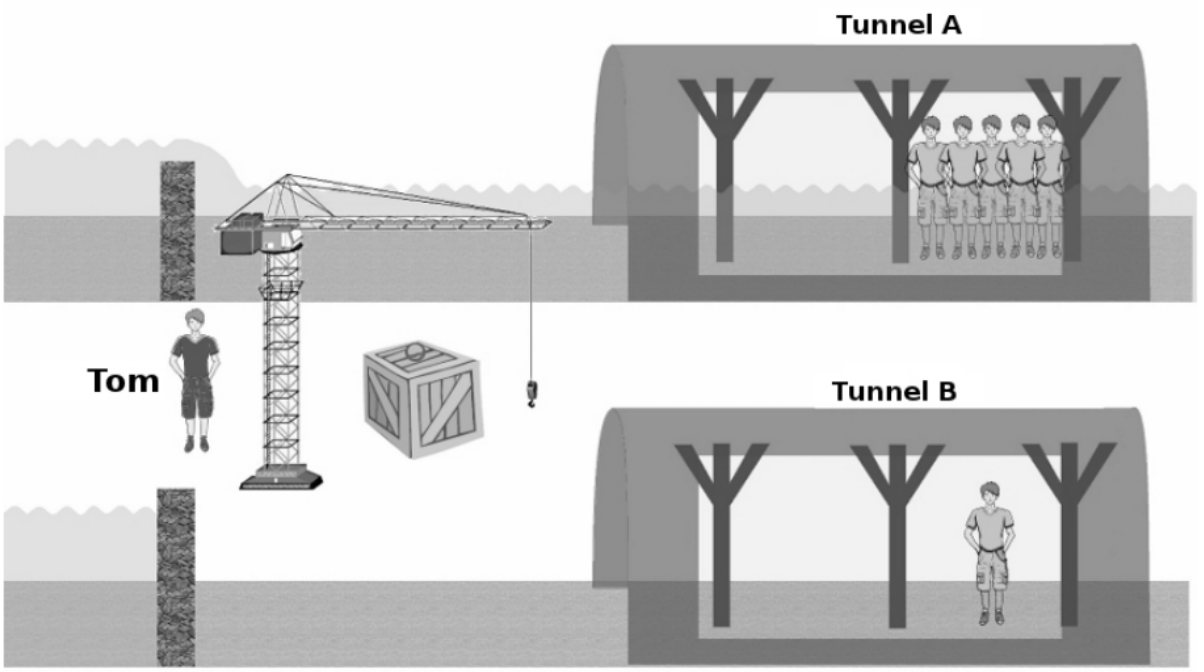
Working hypothesis: attributions of **moral** and **legal responsibility** share a similar cognitive architecture



flooded mine dilemma *(trolley problem variation)*

- Experiments show that people are more prone to blame an agent for an action:

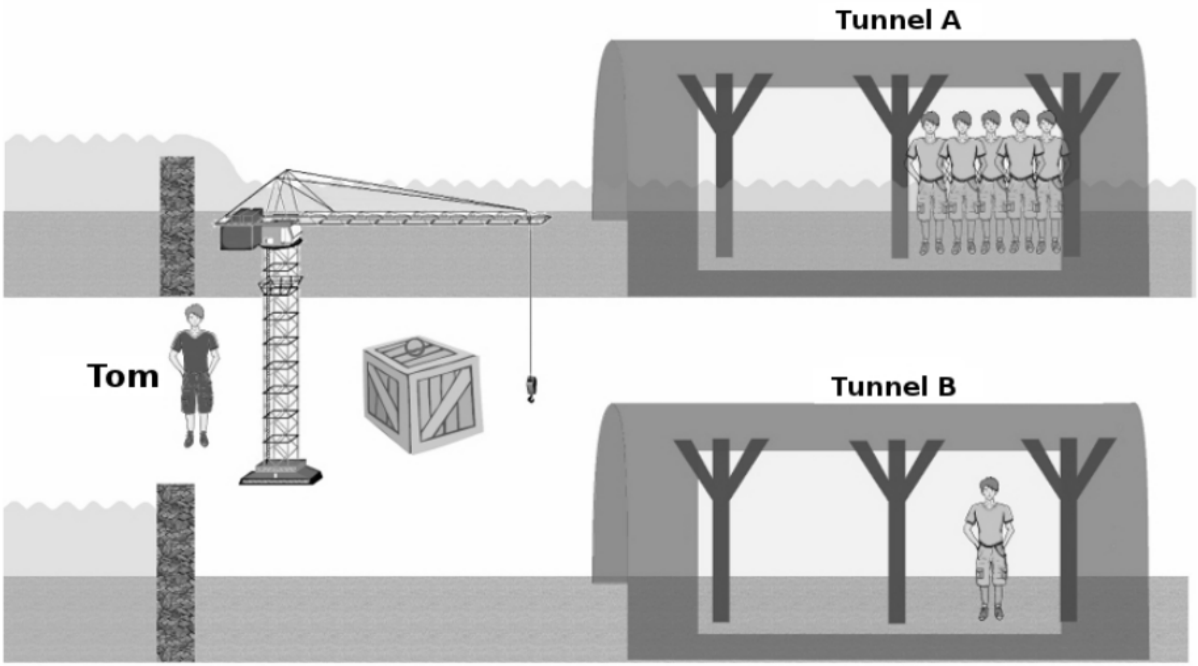
[A. Saillenfest and J.-L. Dessalles. Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. CogSCI 2012, pages 947–952]



flooded mine dilemma *(trolley problem variation)*

- Experiments show that people are more prone to blame an agent for an action:
  - the more the outcome is severe,
  - the more they are closer to the victims,
  - the more the outcome follows the action.

[A. Saillenfest and J.-L. Dessalles. Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. CogSCI 2012, pages 947–952]



flooded mine dilemma *(trolley problem variation)*

- Experiments show that people are more prone to blame an agent for an action:
  - the more the outcome is severe,
  - the more they are closer to the victims,
  - the more the outcome follows the action.
- The cognitive model of ***Simplicity Theory*** predicts these results.

[A. Saillenfest and J.-L. Dessalles. Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. CogSCI 2012]

Simplicity theory

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

Simplicity theory

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness** $U(s) = C_W(s) - C_D(s)$

Simplicity theory

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

causal complexity
concerning how the world generates the situation

Simplicity theory

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.
- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

description complexity

concerning how to identify the situation

causal complexity

concerning how the world generates the situation

Simplicity theory

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

description complexity
concerning how to identify the situation

causal complexity
concerning how the world generates the situation

The two complexities are defined following Kolmogorov complexity.

Kolmogorov complexity

length in bits of the **shortest** program generating a string
description of an object

Kolmogorov complexity

length in bits of the **shortest** program generating a string
description of an object

string		equivalent programs
“2222222222222222222222222”	=	“2” + “2” + … + “2”
	=	“2” * 25
	=	“2” * 5^2

Kolmogorov complexity

length in bits of the **shortest** program generating a string description of an object

string		equivalent programs
“2222222222222222222222222”	=	“2” + “2” + … + “2”
	=	“2” * 25
	=	“2” * 5^2

depends on the available operators!!

Simplicity theory

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

description complexity

about how to identify the situation

length of shortest program determining the situation

causal complexity

about how the world generates the situation

length of shortest program creating the situation

Simplicity theory

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

description complexity

about how to identify the situation

length of shortest program determining the situation

instructions = **mental operators**

causal complexity

about how the world generates the situation

length of shortest program creating the situation

instructions = **causal operators**

Simplicity theory

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

SIMULATION

causal complexity

about how the world generates the situation

length of shortest program creating the situation

instructions = **causal operators**

description complexity

about how to identify the situation

length of shortest program determining the situation

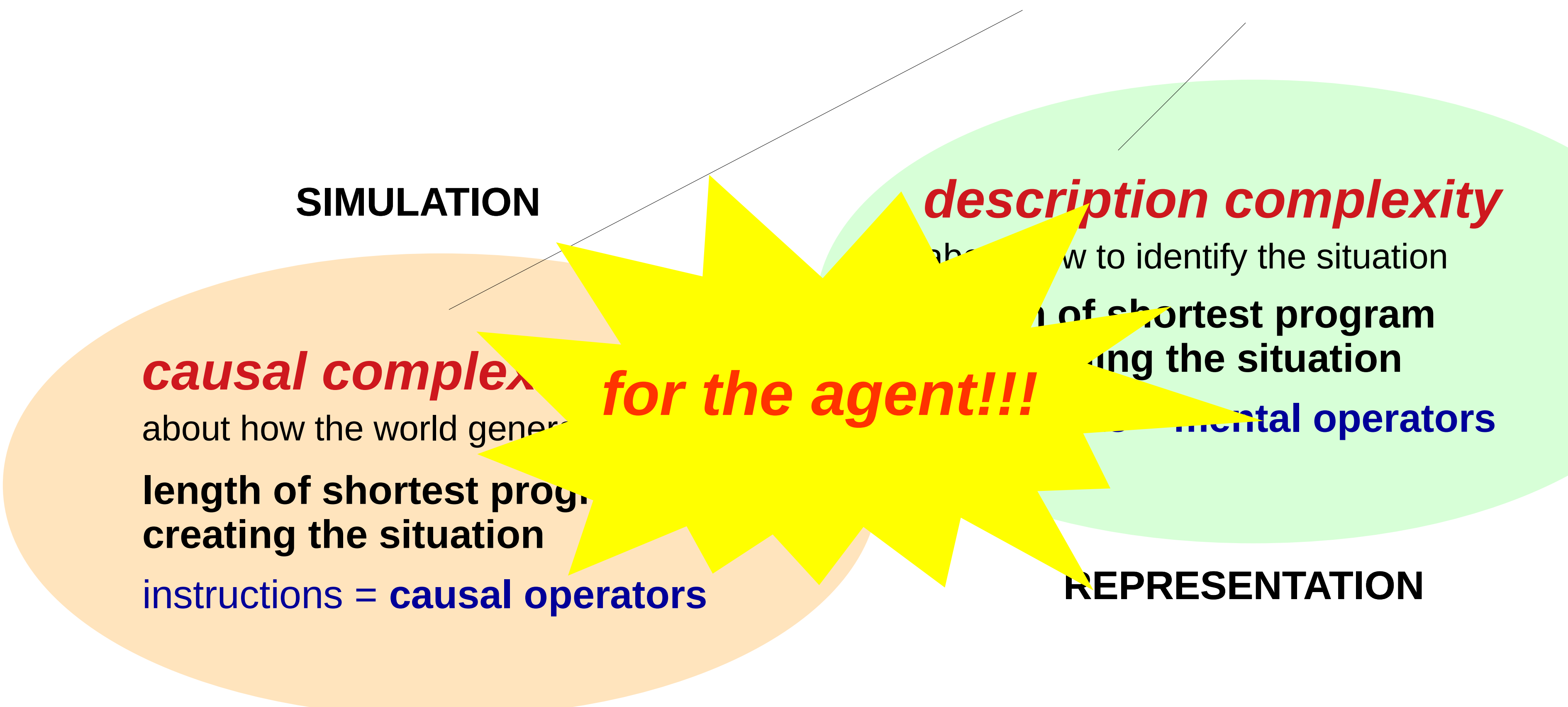
instructions = **mental operators**

REPRESENTATION

Simplicity theory

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness** $U(s) = C_W(s) - C_D(s)$



$$U(s) = C_W(s) - C_D(s)$$

(in a fair extraction)

22222222222222 is more unexpected than **21658367193445**

$$U(s) = C_W(s) - C_D(s)$$

(in a fair extraction)

22222222222222 is more unexpected than **21658367193445**

↓

meeting Obama is more unexpected than **meeting Dupont**

(or any other famous person) (or any other unknown person)

meeting an old of friend of mine

(or any other known person)

Unexpectedness *captures* ***plausibility***

Simplicity Theory: Intention

- Focusing on intensity, we can capture anticipation as:

$$E_h(s) = E(s) - U(s)$$

unexpectedness

emotion

what the situation induces to the agent

reward inverse model

Simplicity Theory: Intention

- Focusing on intensity, we can capture anticipation as:

$$E_h(s) = E(s) - U(s)$$

unexpectedness

emotion

what the situation induces to the agent

reward inverse model

- If the agent *A* expects that the best way to bring about *s* is via *a*:

$$U^A(s) = U^A(a) + U^A(s||a)$$

Simplicity Theory: Intention

- Focusing on intensity, we can capture anticipation as:

$$E_h(s) = E(s) - U(s)$$

emotion
what the situation induces to the agent
reward inverse model

unexpectedness

- If the agent *A* expects that the best way to bring about *s* is via *a*:

$$U^A(s) = U^A(a) + U^A(s||a)$$

$$I(a) = E^A(s) - U^A(s||a) - U^A(a)$$

intention as driven by anticipated emotional effects

Simplicity Theory: Intention

- Focusing on intensity, we can capture anticipation as:

$$E_h(s) = E(s) - U(s)$$

unexpectedness

emotion

what the situation induces to the agent

reward inverse model

- If the agent *A* expects that the best way to bring about *s* is via *a*:

$$U^A(s) = U^A(a) + U^A(s||a)$$

inadvertence

$$I(a) = E^A(s) - U^A(s||a) - U^A(a)$$

intention as driven by anticipated emotional effects

Simplicity Theory: Moral responsibility

- Difference between intention and moral responsibility is one of **point of views**.

computed by A

$$I(a) = E^A(s) - U^A(s||a) - U^A(a)$$

Simplicity Theory: Moral responsibility

- Difference between intention and moral responsibility is one of **point of views**.

computed by A

$$I(a) = E^A(s) - U^A(s||a) - U^A(a)$$

computed by an observer O

computed by a model of A

$$M(a) = E(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a)$$

Simplicity Theory: Moral responsibility

- Difference between intention and moral responsibility is one of **point of views**.

computed by A

$$I(a) = E^A(s) - U^A(s||a) - U^A(a)$$

computed by an observer O

reward inverse model

computed by a model of A

prescribed role, reasonable standard

$$M(a) = E(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a)$$

Simplicity Theory: Moral responsibility

- Difference between intention and moral responsibility is one of **point of views**.

computed by A

$$I(a) = E^A(s) - U^A(s||a) - U^A(a)$$

computed by an observer O

reward inverse model

computed by a model of A

prescribed role, reasonable standard

$$M(a) = E(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a)$$

- Introducing causal responsibility $R^{\downarrow A}(a,s) = C_W(s) - C_W^{\downarrow A}(s||a)$

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a)$$

Simplicity Theory: Moral responsibility

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a)$$

actualized emotion
for observer O
+

causal responsibility
attributed to A
+

conceptual remoteness
for observer O
–

inadvertence
attributed to A
–

Simplicity Theory: Moral responsibility

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a)$$

actualized emotion *for observer O*	causal responsibility *attributed to A*	conceptual remoteness *for observer O*	inadvertence *attributed to A*
+	+	–	–

- From moral to legal responsibility:
  - **equity before the law** (e.g. the "death of a star" case)

Simplicity Theory: Moral responsibility

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a)$$

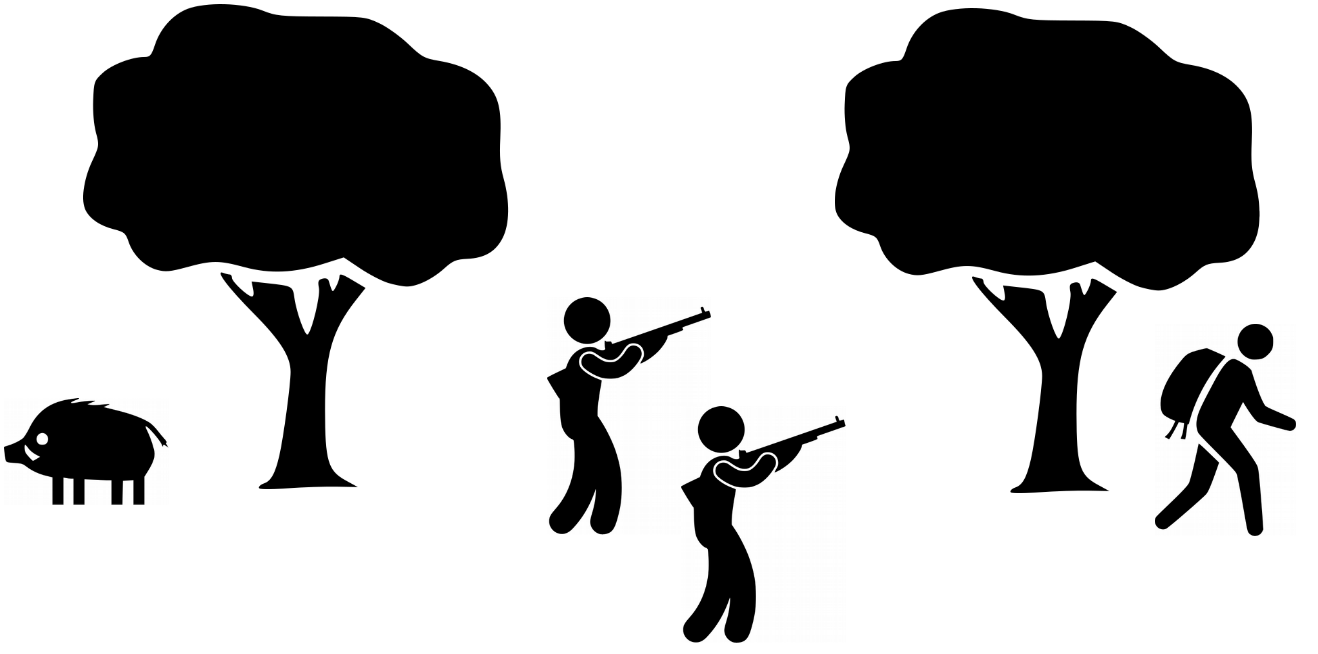
actualized emotion *for observer O*	causal responsibility *attributed to A*	conceptual remoteness *for observer O*	inadvertence *attributed to A*
+	+	–	–

- From moral to legal responsibility:
  - **equity before the law** (e.g. the “death of a star” case)
  - law, as a reward system, defines emotion

Example 1: Negligent hunters

Summers v. Tice (1948), 33 Cal.2d 80, 199 P.2d 1

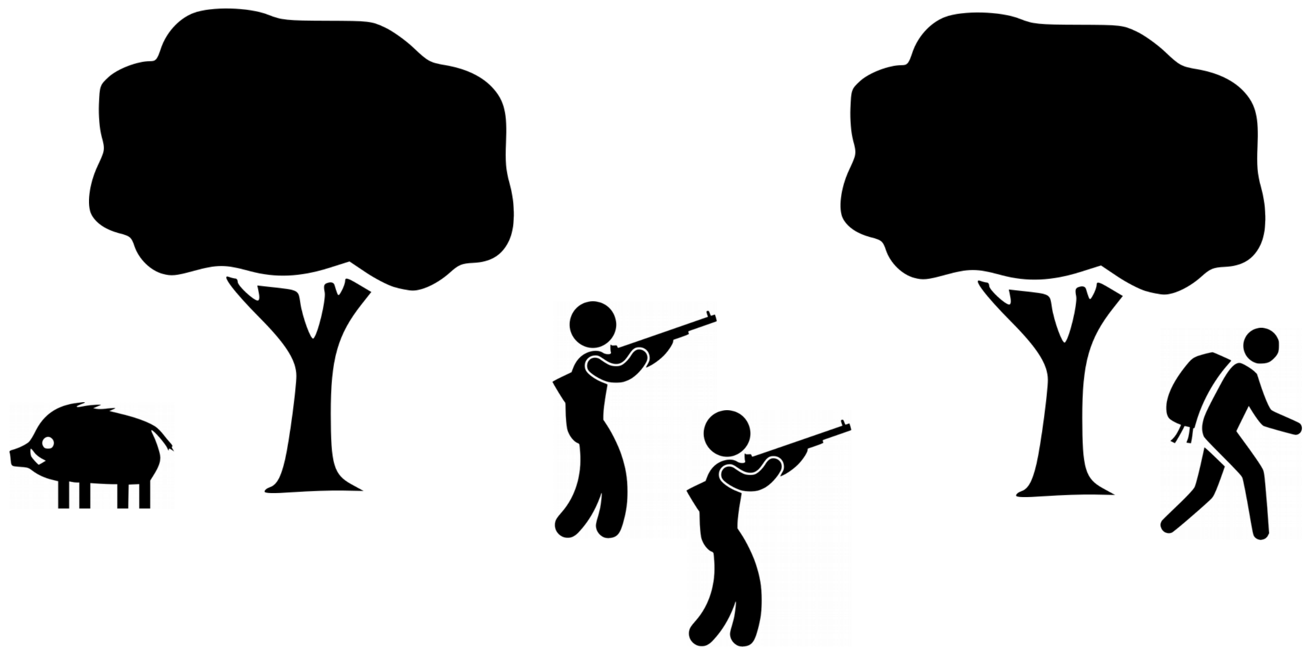
Two hunters shot at the same time harming their guide.



Example 1: Negligent hunters

Summers v. Tice (1948), 33 Cal.2d 80, 199 P.2d 1

Two hunters shot at the same time harming their guide.



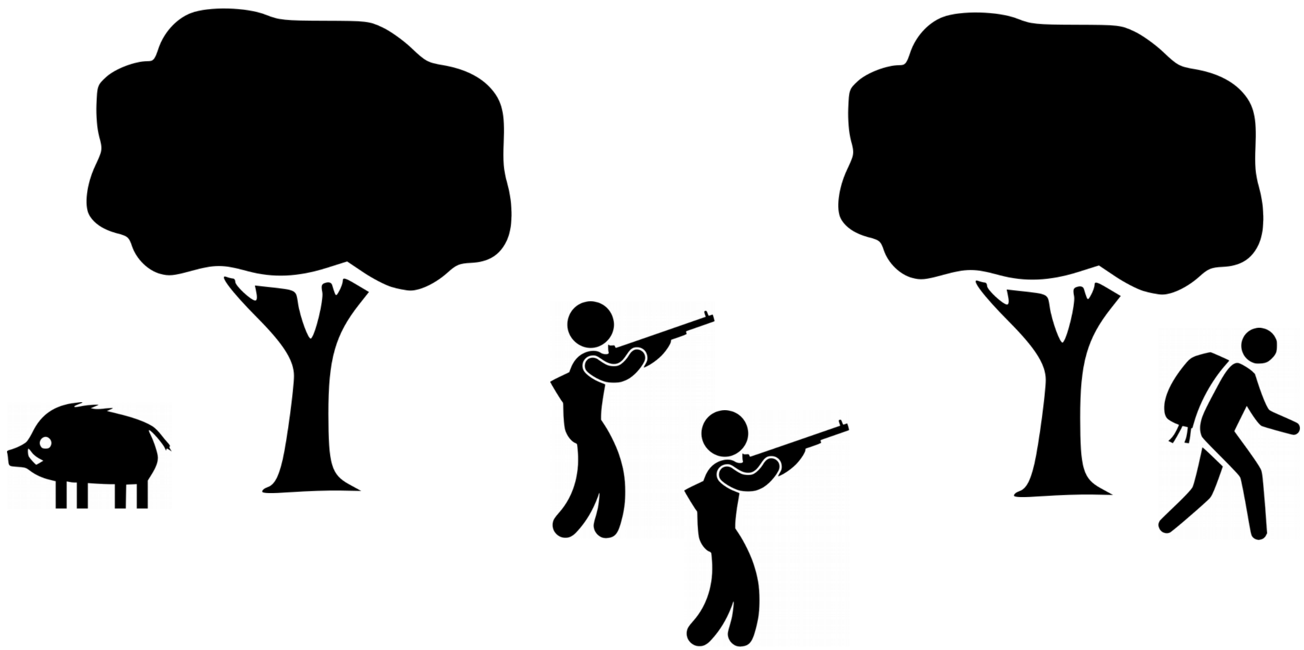
$$C_W^{A_1}(s||a_1) = C_W^{A_2}(s||a_2) \gg 0$$

they thought the harm was impossible

Example 1: Negligent hunters

Summers v. Tice (1948), 33 Cal.2d 80, 199 P.2d 1

Two hunters shot at the same time harming their guide.



$$C_W^{A_1}(s||a_1) = C_W^{A_2}(s||a_2) \gg 0$$

they thought the harm was impossible

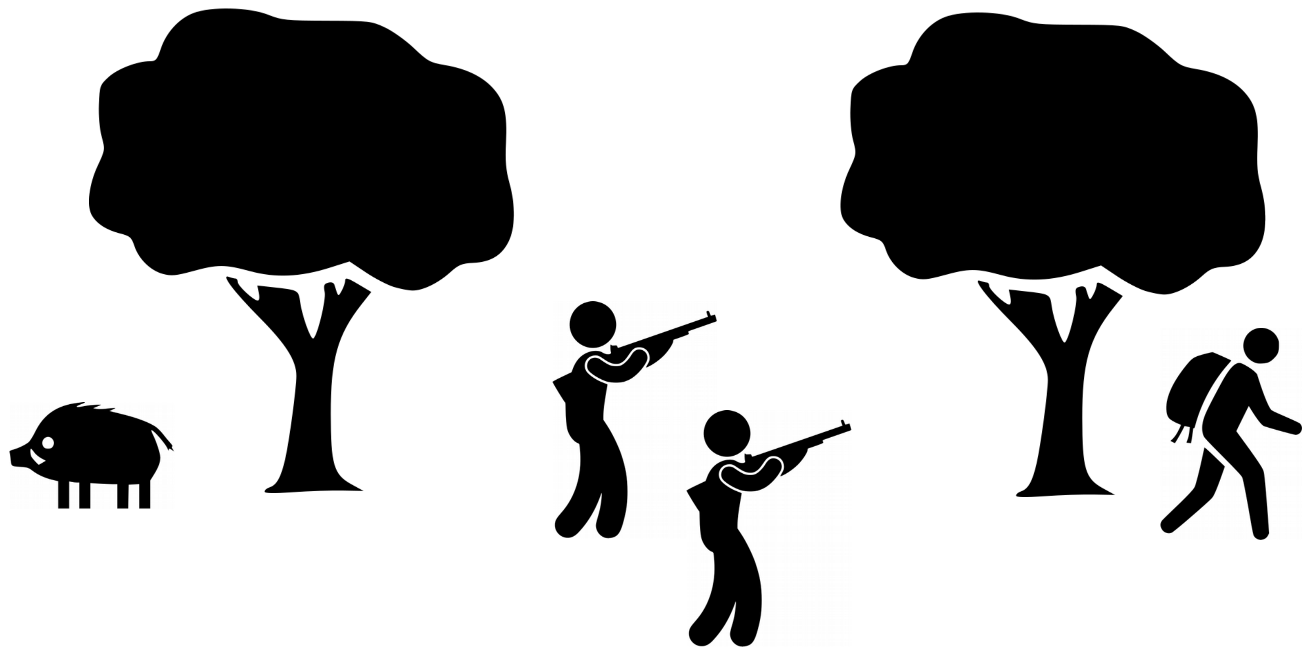
$$C_W^{\downarrow A_1}(s||a_1) = C_W^{\downarrow A_2}(s||a_2) > 0$$

but it was reasonable to consider the danger

Example 1: Negligent hunters

Summers v. Tice (1948), 33 Cal.2d 80, 199 P.2d 1

Two hunters shot at the same time harming their guide.



$C_W^{A_1}(s||a_1) = C_W^{A_2}(s||a_2) \gg 0$ *they thought the harm was impossible*

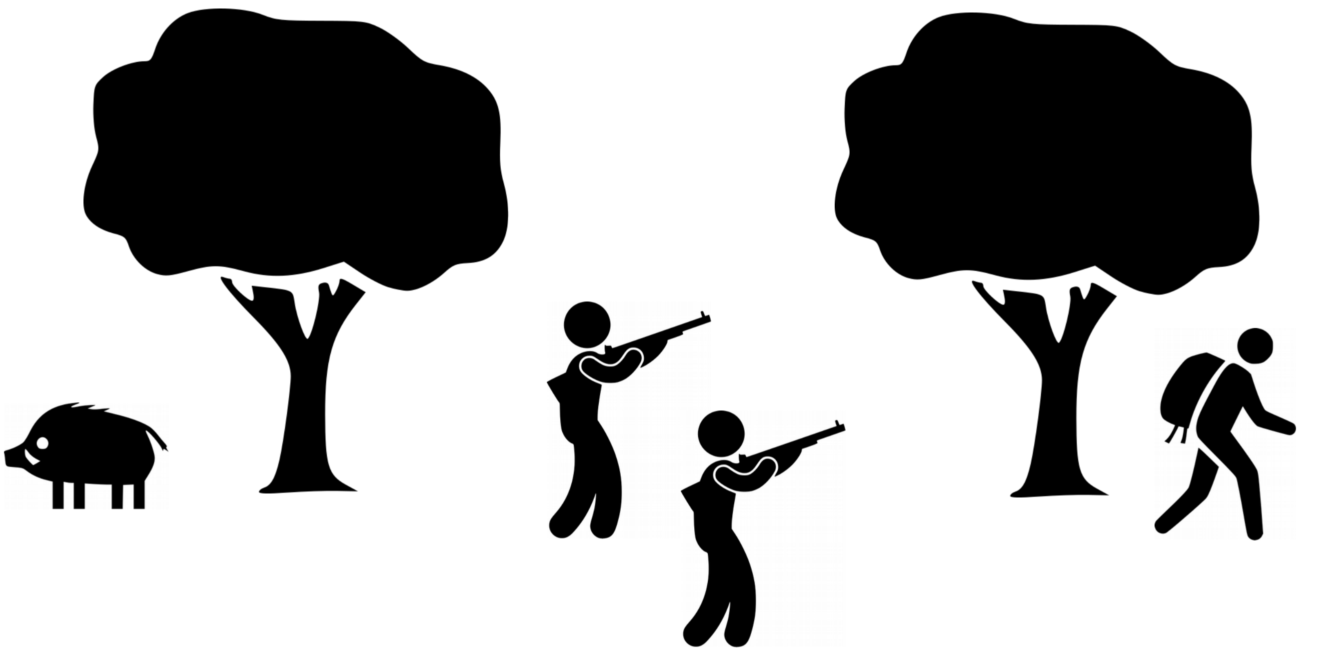
$C_W^{\downarrow A_1}(s||a_1) = C_W^{\downarrow A_2}(s||a_2) > 0$ *but it was reasonable to consider the danger*

$R^{\downarrow A_1}(a_1, s) = R^{\downarrow A_2}(a_2, s) > 0$ *therefore they're* ***(morally) equally responsible.***

Example 1: Negligent hunters

Summers v. Tice (1948), 33 Cal.2d 80, 199 P.2d 1

Two hunters shot at the same time harming their guide.



$C_W^{A_1}(s||a_1) = C_W^{A_2}(s||a_2) \gg 0$ *they thought the harm was impossible*

$C_W^{\downarrow A_1}(s||a_1) = C_W^{\downarrow A_2}(s||a_2) > 0$ *but it was reasonable to consider the danger*

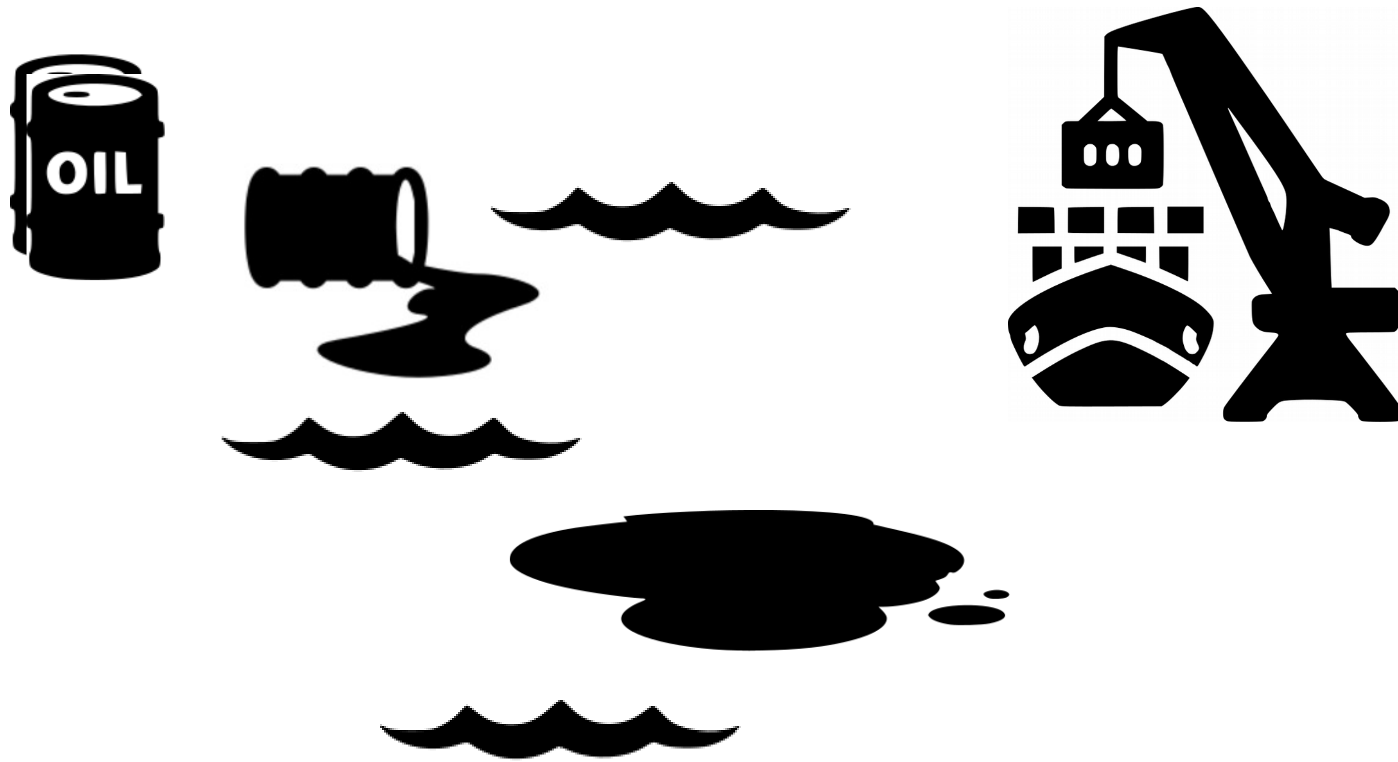
$R^{\downarrow A_1}(a_1, s) = R^{\downarrow A_2}(a_2, s) > 0$ *therefore they're* ***(morally) equally responsible.***

$$N^A(a,s) = C_W^A(s||a) - C_W^{\downarrow A}(s||a)$$ ← ***negligence***

Example 2: Navigating oil

Overseas Tankship (UK) Ltd v. Morts Dock and Eng. Co Ltd – “Wagon Mound (No. 1)” (1961), UKPC 2.

At a landing stage oil was spilled for days in the sea.



Example 2: Navigating oil

Overseas Tankship (UK) Ltd v. Morts Dock and Eng. Co Ltd – “Wagon Mound (No. 1)” (1961), UKPC 2.

At a landing stage oil was spilled for days in the sea.

It was then ignited during works on a ship nearby.

Example 2: Navigating oil

Overseas Tankship (UK) Ltd v. Morts Dock and Eng. Co Ltd – “Wagon Mound (No. 1)” (1961), UKPC 2.

At a landing stage oil was spilled for days in the sea.

It was then ignited during works on a ship nearby.

$C_W^{\downarrow A}(s_1||a) \sim 0$ with poor maintenance, sea contamination by oil leakage predictable

Example 2: Navigating oil

Overseas Tankship (UK) Ltd v. Morts Dock and Eng. Co Ltd – “Wagon Mound (No. 1)” (1961), UKPC 2.

At a landing stage oil was spilled for days in the sea.

It was then ignited during works on a ship nearby.

$C_W^{\downarrow A}(s_1||a) \sim 0$ with poor maintenance, sea contamination by oil leakage predictable

$C_W^{\downarrow A}(s_2||s_1) \gg 0$ fire after oil leakage in sea difficult to occur

Example 2: Navigating oil

Overseas Tankship (UK) Ltd v. Morts Dock and Eng. Co Ltd – “Wagon Mound (No. 1)” (1961), UKPC 2.

At a landing stage oil was spilled for days in the sea.

It was then ignited during works on a ship nearby.

$C_W^{\downarrow A}(s_1||a) \sim 0$ with poor maintenance, sea contamination by oil leakage predictable

$C_W^{\downarrow A}(s_2||s_1) \gg 0$ fire after oil leakage in sea difficult to occur

$R^{\downarrow A}(a, s_2) \sim 0$ therefore, defendant is not responsible

Example 2: Navigating oil

Overseas Tankship (UK) Ltd v. Morts Dock and Eng. Co Ltd – “Wagon Mound (No. 1)” (1961), UKPC 2.

At a landing stage oil was spilled for days in the sea.

It was then ignited during works on a ship nearby.

$C_W^{\downarrow A}(s_1||a) \sim 0$ with poor maintenance, sea contamination by oil leakage predictable

$C_W^{\downarrow A}(s_2||s_1) \gg 0$ fire after oil leakage in sea difficult to occur

$R^{\downarrow A}(a, s_2) \sim 0$ therefore, defendant is not responsible

$$F^A(a,s) = -U^{\downarrow A}(s||a)$$ ← ***foreseeability***

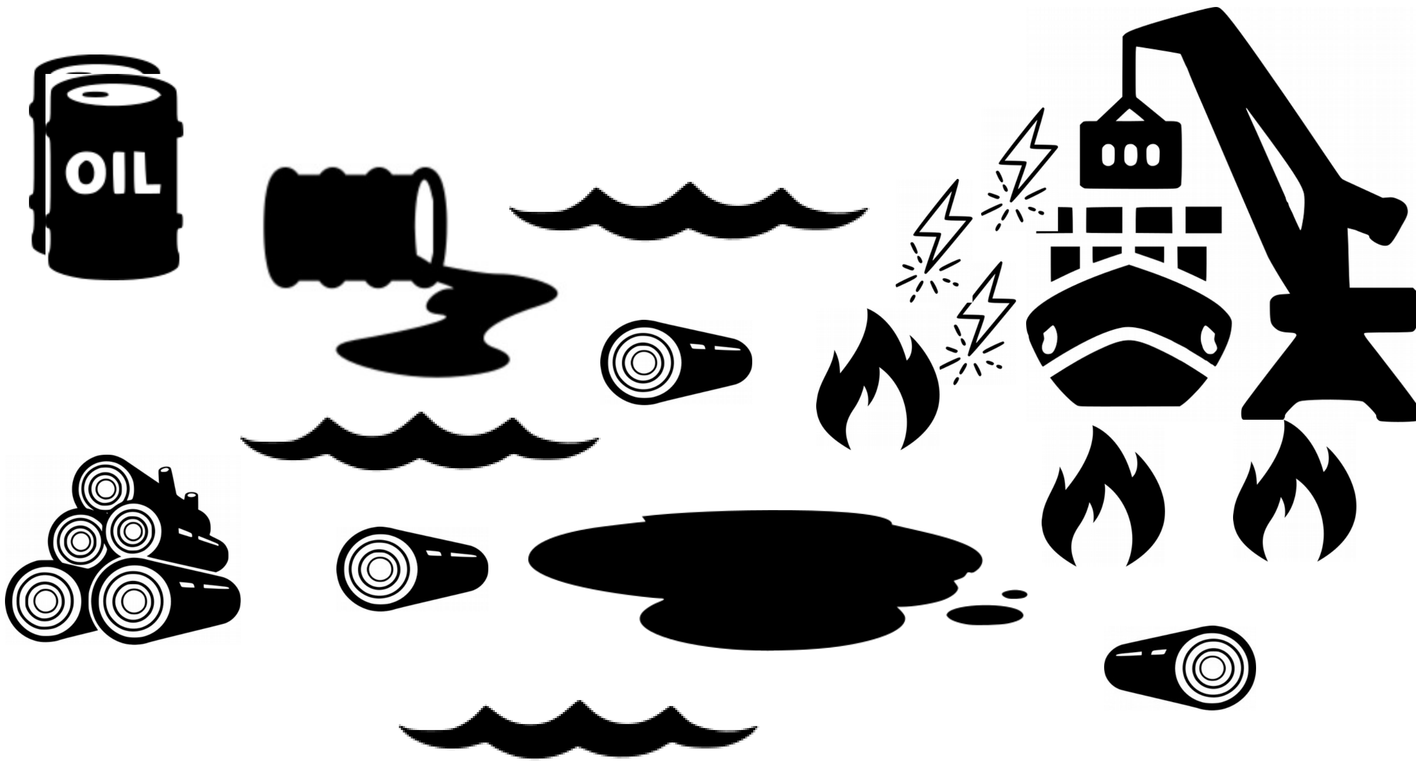
Example 3: Navigating oil, *continued*

Overseas Tankship (UK) Ltd v The Miller Steamship Co – "Wagon Mound (No. 2)" (1967), 1 AC 617.

At a landing stage oil was spilled for days in the sea.

It was then ignited during works on a ship nearby.

NEW EVIDENCE: flammable objects in the water.



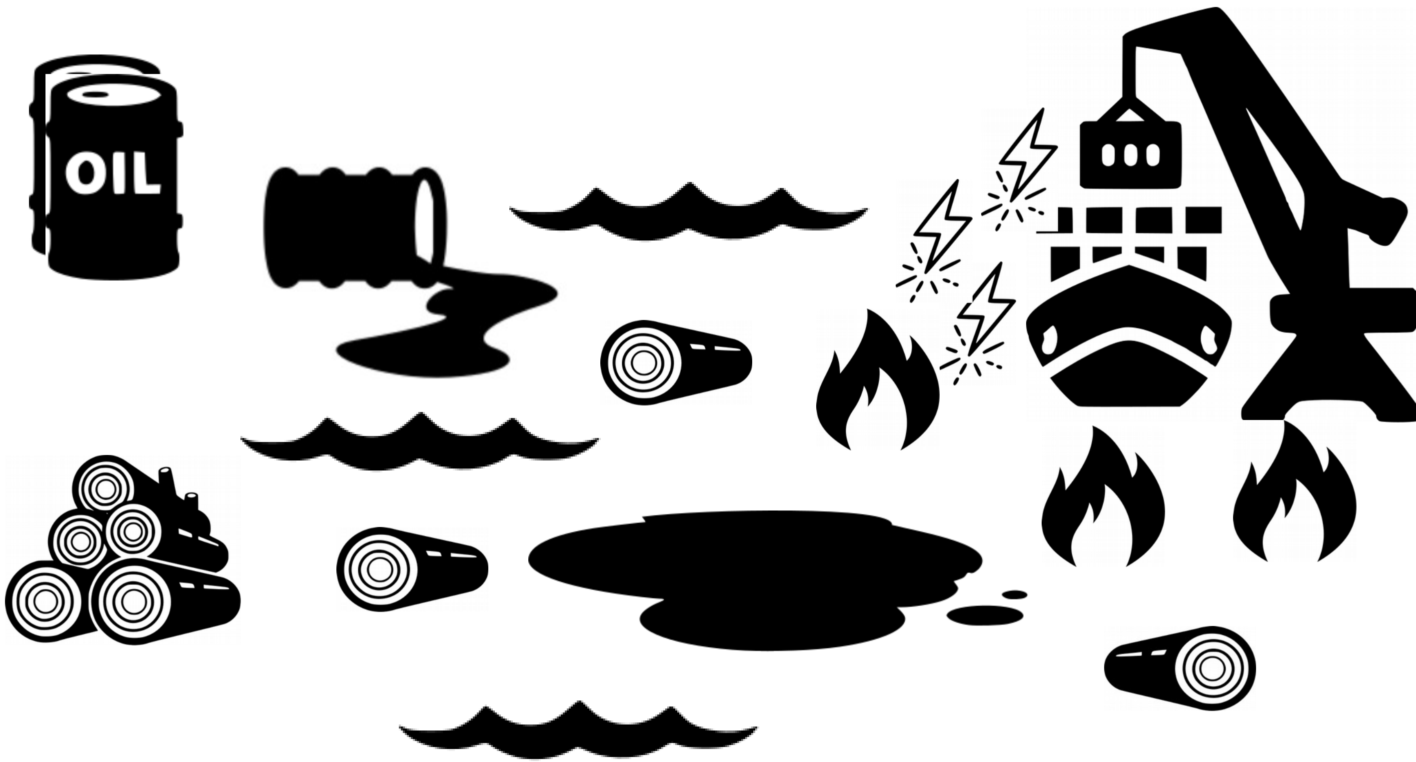
Example 3: Navigating oil, *continued*

Overseas Tankship (UK) Ltd v The Miller Steamship Co – “Wagon Mound (No. 2)” (1967), 1 AC 617.

At a landing stage oil was spilled for days in the sea.

It was then ignited during works on a ship nearby.

NEW EVIDENCE: flammable objects in the water.



1st argument: foreseeability

$C_W^{\downarrow A}(s_1||a) \sim 0$ with poor maintenance, sea contamination by oil leakage predictable

$C_W^{\downarrow A}(s_2||s_1) > 0$ fire after oil leakage **possible**, ***because of flammable objects***

$R^{\downarrow A}(a, s_2) > 0$ therefore, defendant **is** responsible

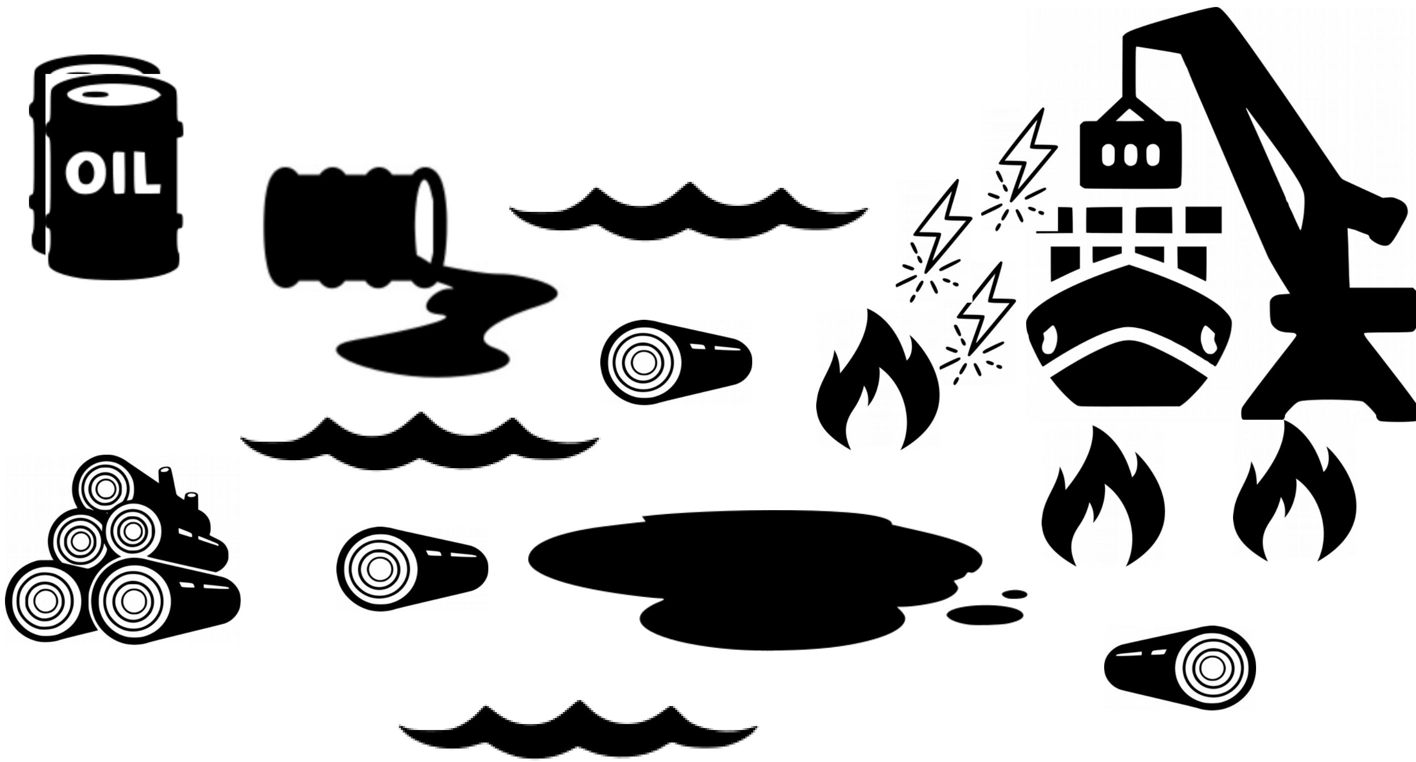
Example 3: Navigating oil, *continued*

Overseas Tankship (UK) Ltd v The Miller Steamship Co – "Wagon Mound (No. 2)" (1967), 1 AC 617.

At a landing stage oil was spilled for days in the sea.

It was then ignited during works on a ship nearby.

NEW EVIDENCE: flammable objects in the water.



2nd argument: weighting of risks (anticipations)

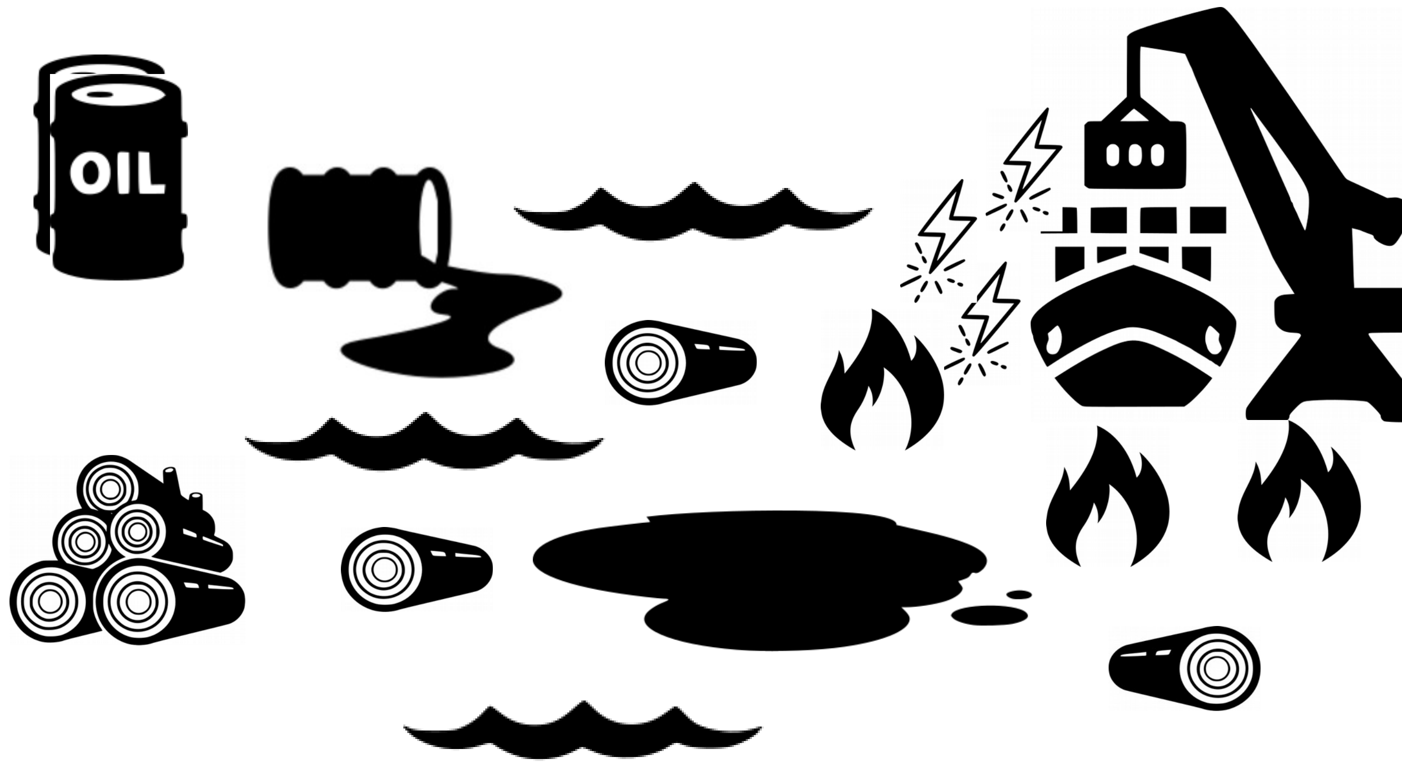
Example 3: Navigating oil, *continued*

Overseas Tankship (UK) Ltd v The Miller Steamship Co – “Wagon Mound (No. 2)” (1967), 1 AC 617.

At a landing stage oil was spilled for days in the sea.

It was then ignited during works on a ship nearby.

NEW EVIDENCE: flammable objects in the water.



2nd argument: weighting of risks (anticipations)

$$M(a) = \underbrace{E(s)} - U^{\downarrow A}(s||a) - U^{\downarrow A}(a)$$

risk ⟶ $$K^{A}(a,s) = E(s) - U^{\downarrow A}(s||a) = E(s) + F^{A}(a,s)$$

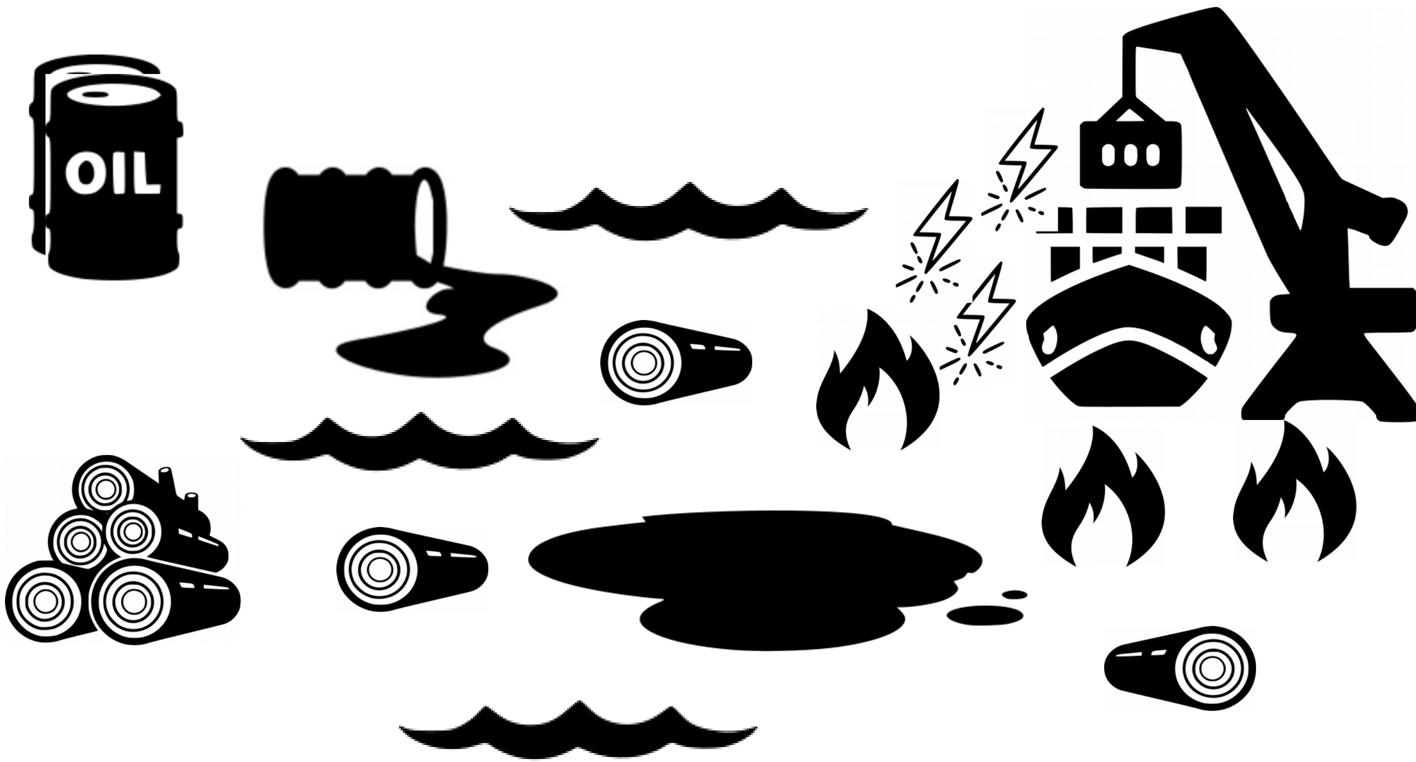
Example 3: Navigating oil, *continued*

Overseas Tankship (UK) Ltd v The Miller Steamship Co – "Wagon Mound (No. 2)" (1967), 1 AC 617.

At a landing stage oil was spilled for days in the sea.

It was then ignited during works on a ship nearby.

NEW EVIDENCE: flammable objects in the water.



2nd argument: weighting of risks (anticipations)

$$M(a) = E(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a)$$

risk →

$$K^{A}(a,s) = E(s) - U^{\downarrow A}(s||a) = E(s) + F^{A}(a,s)$$

risk as generalization of foreseeability: Hart and Honoré's view!

Conclusions

- Our contribution attempts to open an alternative research track for the computation of responsibility in AI & Law.

Conclusions

- Our contribution attempts to open an alternative research track for the computation of responsibility in AI & Law.
- Underlying model ***derived*** from general cognitive functions (**SIMULATION, REPRESENTATION, REWARD INVERSE MODEL**)

Conclusions

- Our contribution attempts to open an alternative research track for the computation of responsibility in AI & Law.
- Underlying model ***derived*** from general cognitive functions (**SIMULATION, REPRESENTATION, REWARD INVERSE MODEL**)
- It enables a smoother transition from moral to legal reasoning, and provides grounds to quantify legal concepts.

Conclusions

- Our contribution attempts to open an alternative research track for the computation of responsibility in AI & Law.
- Underlying model ***derived*** from general cognitive functions (**SIMULATION, REPRESENTATION, REWARD INVERSE MODEL**)
- It enables a smoother transition from moral to legal reasoning, and provides grounds to quantify legal concepts.
- Computation integrates quantitative and structural aspects: potential ground for unifying other approaches, e.g. exploiting explicit knowledge and probabilistic information.
  - further work is needed for a complete operationalization and for detailed comparisons