# Logic-based Formulation of Ethical Principles

John Hooker
*Carnegie Mellon University*

ICLP 2021

Some results represent joint work with

Thomas Donaldson
*University of Pennsylvania*

Tae Wan Kim
*CMU*

# Ethics in AI

- **There is rapidly growing interest in AI ethics**
  - *Mainly to avoid **bias** in AI-based decisions.*
  - *But also to incorporate **general ethical principles** into AI systems.*
    - "Value alignment"

# Ethics in AI

- **There is rapidly growing interest in AI ethics**
  - *Mainly to avoid **bias** in AI-based decisions.*
  - *But also to incorporate **general ethical principles** into AI systems.*
    - "Value alignment"
- **Our goals:**
  - *Show that principles can be stated **rigorously** enough to allow **logic-based formulation**.*
    - This requires some background in **deontological ethics**.
  - *Show that logic-based formulation enables **value alignment** to incorporate the ethical principles.*

# Basic assumptions

- **Acting for reasons**
  - *Freely chosen action is based on a rationale.*
- **Universality of reason**
  - *Justification is independent of the reasoner.*

.

# Basic assumptions

- **Acting for reasons**
  - *Freely chosen action is based on a rationale.*
- **Universality of reason**
  - *Justification is independent of the reasoner.*
- We **deduce** ethical principles from these assumptions.
  - *This is the **deontological** approach to ethics.*
  - ***Deontology** = What is required.*
    - Ethical principles represent what is required for the possibility of free action.

# Acting for reasons

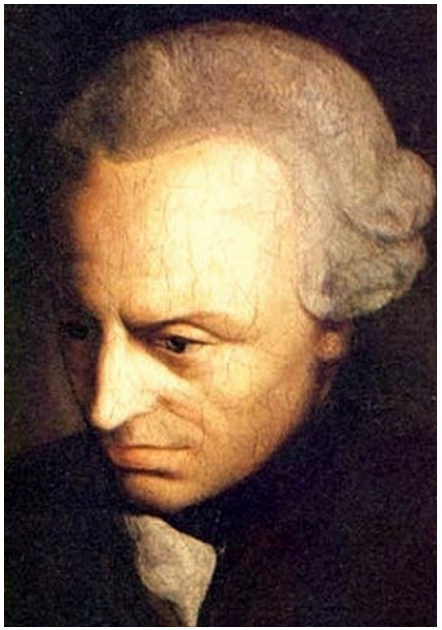
- Basic premise:  We always act for a reason.
  - *Every action has a rationale.*
- Why?
  - *This is how we distinguish **freely chosen action** from mere behavior.*
    - An MRI machine can detect our decisions **before we make them**.
    - If decisions are determined by **biological causes**, how can they be freely chosen?

# Acting for reasons

- Solution:
  - *Freely chosen actions have **two kinds of explanation**:*
    - A biological cause
    - A rationale provided by the agent
  - *For example:*
    - A hiccup has **only** a biological explanation.
      Not a freely chosen action.
    - Drinking water to stop hiccups has
      **2 explanations**: a biological cause
      and a rationale.  A freely chosen action.

# Acting for reasons

- Dual standpoint theory
  - *Originally proposed by Immanuel Kant.*
    - *Grundlegung zur Metaphysik der Sitten* (1785)
    - Recent versions: *Nagel (1986), Korsgaard (1996), Nelkin (2000), Bilgrami (2006).*
  - *Provides a **basis for ethics**.*
    - Ethical principles are **necessary conditions** for the logical coherence of an action's rationale.



# Universality of reason

- What is rational **does not depend on who I am**.
  - *I don't get to have my own logic.*
  - *In particular, if I view a reason as justifying an action for me, I must view it as justifying the same action* ***for anyone to whom the reason applies****.*
- The assumption underlies science and all forms of rational inquiry.
  - *Ethics assumes nothing more.*

# Principles

- We sketch **deontological arguments** for three ethical principles.
    - Based on assumptions just stated.
  - ***Generalization principle***
  - ***Autonomy principle***
  - ***Utilitarian principle***
- We show how to express the principles in **quantified modal logic**.
  - *To allow application to value alignment.*

# Generalization principle

- **Example**
- Suppose I steal a watch from a shop.
- I have 2 reasons:
  - *I want a new watch.*
  - *I won't get caught.*
    - Security at the shop is lax.

# Generalization principle

- **Example**
- Suppose I steal a watch from a shop.
- I have 2 reasons:
  - *I want a new watch.*
  - *I won't get caught.*
    - Security at the shop is lax.

- These are not psychological causes or motivations.
  - *They are consciously adduced reasons for the theft.*
    - There may be other reasons, of course.

# Example - Theft

- Due to universality of reason, I am making a decision for everyone:
  - *All who want a watch and think they won't get caught should steal one.*

# Example - Theft

- Due to universality of reason, I am making a decision for everyone:
  - *All who want a watch and think they won't get caught should steal one.*
- But I know that if all do this, they will get caught.
  - *The shop will install security.*
  - *My reasons will no longer apply to **me**.*

# Example - Theft

- Due to universality of reason, I am making a decision for everyone:
  - *All who want a watch and think they won't get caught should steal one.*
- But I know that if all do this, they will get caught.
  - *The shop will install security.*
  - *My reasons will no longer apply to **me**.*
- I am not saying that all these people actually **will** steal watches.
  - *Only that if they did, my reasons would no longer apply*.

# Example - Theft

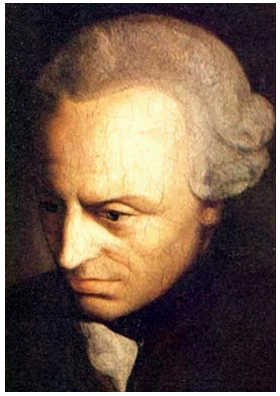
- My reasons are **inconsistent** with the assumption that people will act on them.
- I am caught in a contradiction.
  - *I am deciding that these reasons justify theft for **me**.*
  - *But I am **not** deciding that these reasons justify theft for **others**.*
  - *I can't have it both ways.*

# Example - Theft

- My reasons are **inconsistent** with the assumption that people will act on them.
- I am caught in a contradiction.
  - *I am deciding that these reasons justify theft for **me**.*
  - *But I am **not** deciding that these reasons justify theft for **others**.*
  - *I can't have it both ways.*
- More generally…
  - *Universal theft merely for personal benefit would **undermine the institution of property**.*
    - Purpose of theft is to benefit from property rights.

# Generalization principle

- It should be **rational** for me to believe that the **reasons** for my action are **consistent** with the assumption that **everyone to whom the same reasons apply acts the same way**.
  - *Historically inspired by Kant's Categorical Imperative, but different and more precise.*
  - *Takes "rationality" as a primitive and unexplained notion, but this is true to some extent of all science.*



# Example - Cheating

- What is wrong with cheating on an exam?
- My reasons:
  - *I will get a better grade and therefore a better job.*
  - *I can get away with it.*
- I know that these reasons apply to nearly all students.
  - *If they act accordingly, grades will be meaningless, or exams strictly proctored.*
  - *This defeats one or both of my reasons.*
  - *So, cheating for these reasons **violates** the generalization principle.*

# Example - Agreements

- Breaking an agreement normally violates the generalization principle.
- Reason:
  - *Convenience or profit.*
- These reasons apply to most agreements
  - *If agreements were broken for mere convenience, it would be impossible to **make** agreements.*
  - *And therefore impossible to **achieve one's purposes** by **breaking** them.*
  - *The whole point of having an agreement is that you keep it when **you don't want to keep it**.*

# Example - Lying

- Lying for mere convenience violates the generalization principle.
  - *…if the reason for lying assumes that people will believe the lie.*
  - *If everyone lied when convenient, no one would believe the lies.*
    - The possibility of **communication** presupposes a certain amount of credibility.

# Example - Lying

- Lying can be generalizable, depending on the reasons.
- Popular "counterexample"
  - Similar to one posed in Kant's day.
  - *Workers in an Amsterdam office building lied to Nazi police, to conceal whereabouts of Anne Frank and family.*
  - *This is **generalizable**.*
    - If everyone lied for this reason, it would still accomplish the purpose, perhaps even more effectively.
    - There is no need for police to believe the lies.

# Scope of the rationale

- Scope = an agent's necessary and jointly sufficient conditions for performing an act.
  - *An ambulance driver uses the siren, but with no patient.*
  - *His reasons:*
    - He is late picking up his kids at day care, because he misplaced his car keys.
    - The siren will allow him to arrive on time.
    - He can get away with it.
  - *This is generalizable*
    - These reasons seldom apply to an ambulance driver.

# Scope of the rationale

- Scope = an agent's necessary and jointly sufficient conditions for performing an act.
  - *An ambulance driver uses the siren, but with no patient.*
  - *His reasons:*
    - He is late picking up his kids at day care, because he misplaced his car keys.
    - The siren will allow him to arrive on time.
    - He can get away with it.
  - *This is generalizable*
    - These reasons seldom apply to an ambulance driver.
  - *But the scope is too narrow*
    - The details are not necessary.
    - The real reason is that it is important to be on time.

# Action plans

- Since actions always have a rationale, we treat them as **action plans**.
  - *If X, then do Y.*
  - *For example,*
    - **If** I would like to have an item on display in a shop,
      **and** I can get away with stealing it,
      **then** I will steal it.
- An **agent** is a bundle of action plans.
  - *…that are executed when the antecedents are satisfied.*

# Logical formulation

- The first step is to formulate action plans as conditionals.

$C_1(a) =$ Agent $a$ wants an item on display in a shop.
$C_2(a) =$ Agent $a$ can get away with stealing the item.
$A(a) \;\; =$ Agent $a$ will steal the item.

*The action plan is:* $\big(C_1(a) \wedge C_2(a)\big) \Rightarrow_a A(a)$

$\Rightarrow_a$ is not logical entailment but indicates that agent $a$ regards $C_1(a)$ and $C_2(a)$ as justifying $A(a)$.

# Logical formulation

- Modal operators.

$\Box_a S$ = agent $a$ must assent to $S$ to be rational
$\Diamond_a S$ = agent $a$ can be rational in assenting to $S$

Thus $\Diamond_a S \equiv \neg\Box_a\neg S$, as usual.

We will also say
$\Box_a S$ = agent $a$ is rationally constrained to believe $S$
$\Diamond_a S$ = agent $a$ can rationally believe $S$

*The operators have different interpretations than in traditional alethic, epistemic and doxastic logics.*

Note that we don't have $\Box_a S \rightarrow S$

# Logical formulation

- Possibility predicate

$$P(S) = S \text{ is possible}$$

*Possibility is **not** a modal operator here.*

*We can regard this as physical (as opposed to logical) possibility.*

*It is not essential to be more precise at this point.*

# Logical formulation

- Let $C(a) \Rightarrow_a A(a)$ be an action plan

  where $C(a)$ is a conjunction of $a$'s reasons for $A(a)$

- The **generalization principle** is

$$\Diamond_a P\Big(\forall x\big(C(x) \rightarrow A(x)\big) \wedge C(a) \wedge A(a)\Big)$$

*Agent a can rationally believe that it is possible to take action A when reasons C apply, and when all agents to whom reasons C apply take action A.*

# Autonomy

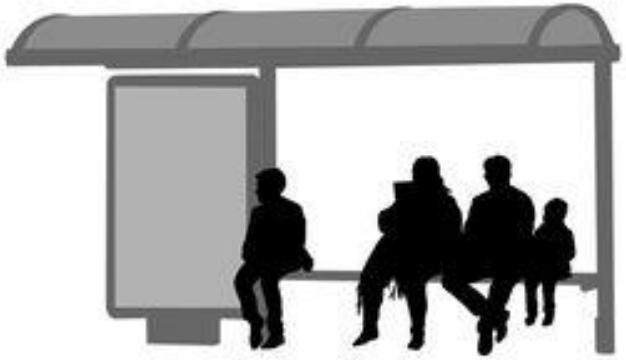
- There is a fundamental obligation to respect autonomy.
  - *This rules out murder, most coercion, slavery, etc.*
  - *But autonomy must be carefully defined.*

# Autonomy

- There is a fundamental obligation to respect autonomy.
  - *This rules out murder, most coercion, slavery, etc.*
  - *But autonomy must be carefully defined.*
- Autonomy is more than "self-law."
  - *I act **autonomously** when I freely make up my own mind about what to do, based on **coherent reasons** I give for my decision*
    - An **agent** is a being that can act autonomously (sometimes called a "moral agent").
    - Today's "autonomous cars" are **not** autonomous.

# Violation of autonomy

- Coercion violates autonomy if it **interferes with an ethical action plan**.
  - *Example.*
    - Action plan: "If I want to catch a bus, and the bus stop is across the street, and no cars are coming, the I will cross the street."
    - If you pull me off the street when no cars are coming, this is a **violation** of my autonomy.
    - If you pull me out of the path of a car I fail to see, this is **coercion** but **no violation** of autonomy.



# Autonomy principle

- My action plan is unethical if I am **rationally constrained to believe it interferes** with the **ethical action plan** of some other agent.

# Autonomy principle

- I must be **rationally constrained** to believe there is a conflict of action plans.
  - *That is, it is **irrational** not to believe this.*
    - If someone falls into a maintenance hole I leave uncovered, this is **not** a violation of autonomy.
    - It is only possible/probable that someone will fall in (a violation of the **utilitarian principle**).

# Autonomy principle

- I must be **rationally constrained** to believe there is a conflict of action plans.
  - *That is, it is **irrational** not to believe this.*
    - If someone falls into a maintenance hole I leave uncovered, this is **not** a violation of autonomy.
    - It is only possible/probable that someone will fall in (a violation of the **utilitarian principle**).
    - But suppose it has a cover that will **collapse** when someone steps on it and is on 5th Ave NYC (a booby trap).
    - I am **rationally constrained** to believe **someone** will fall in.
    - I **violate autonomy**.

# Autonomy principle

- Coercion does not violate autonomy if there is **informed consent**.
  - *Suppose I attend a concert with strict rules against recording the performance.*
    - Ushers compel me to leave when I record it anyway.
    - This is **coercion** but **no violation of my autonomy**.
    - I gave **informed consent** to this possibility.
    - The consent is part of the **antecedent** of my action plan.
    - "If I want to record the performance and am not kicked out for doing so, then I will record it."

# Autonomy principle

- Interference with an **unethical** action plan is **not** a violation of autonomy.
  - *An unethical action plan is not a freely chosen action, because it has no coherent rationale.*
  - *There is **no denial of agency**.*
    - You can defend yourself, because an attack on you is unethical.

# Autonomy principle

- Interference with an **unethical** action plan is **not** a violation of autonomy.
  - *An unethical action plan is not a freely chosen action, because it has no coherent rationale.*
  - *There is **no denial of agency**.*
    - You can defend yourself, because an attack on you is unethical.
  - *Is this a circular reference to "unethical"?*
    - We define "unethical" **recursively**.
    - An action plan is unethical if it violates the **generalization** of **utilitarian** principle, **or** interferes with an ethical action plan.

# Logical formulation

Agent $a$'s action plan $C(a) \Rightarrow_a A(a)$ interferes with agent $b$'s action plan $C'(b) \Rightarrow_b A'(b)$ when

$$\Box_a \neg P\big(A(a) \wedge A'(b)\big) \wedge \Diamond_a P\big(C(a) \wedge C'(b)\big)$$

*Agent a is rationally constrained to believe that the two actions are incompatible,*
*and can rationally believe that that the reasons for the two actions can both apply.*

# Logical formulation

- Example

$C_1(b)$ = agent $b$ wants to catch a bus
$C_2(b)$ = there is a bus stop across the street from $b$
$C_3(b)$ = cars are approaching $b$
$C_4(b)$ = agent $b$ is about to cross the street
$A_1(b)$ = agent $b$ will cross the street
$A_2(a,b)$ = agent $a$ will pull $b$ off the street

*No cars coming*

Agent $a$'s plan: $\left(\boxed{\neg C_3(b)} \wedge C_4(b)\right) \Rightarrow_a A_2(a,b)$

Agent $b$'s plan: $\left(C_1(b) \wedge C_2(b) \wedge \neg C_3(b)\right) \Rightarrow_b A_1(b)$

Agent a's plan interferes with agent $b$'s plan:

*True due to coercion*

$$\boxed{\Box_a \neg P\left(A_1(b) \wedge A_2(a,b)\right)} \wedge$$
$$\Diamond_a P\left(C_1(b) \wedge C_2(b) \wedge \neg C_3(b) \wedge C_4(b)\right)$$

# Logical formulation

- Example

$C_1(b)$ = agent $b$ wants to catch a bus
$C_2(b)$ = there is a bus stop across the street from $b$
$C_3(b)$ = cars are approaching $b$
$C_4(b)$ = agent $b$ is about to cross the street
$A_1(b)$ = agent $b$ will cross the street
$A_2(a,b)$ = agent $a$ will pull $b$ off the street

*No cars coming*

Agent $a$'s plan: $\Big(\boxed{\neg C_3(b)} \wedge C_4(b)\Big) \Rightarrow_a A_2(a,b)$

Agent $b$'s plan: $\Big(C_1(b) \wedge C_2(b) \wedge \neg C_3(b)\Big) \Rightarrow_b A_1(b)$

Agent a's plan interferes with agent $b$'s plan:

$\Box_a \neg P\Big(A_1(b) \wedge A_2(a,b)\Big) \wedge$ *True due to mutually consistent reasons*

$$\boxed{\Diamond_a P\Big(C_1(b) \wedge C_2(b) \wedge \neg C_3(b) \wedge C_4(b)\Big)}$$

# Logical formulation

- Example

$C_1(b)$ = agent $b$ wants to catch a bus
$C_2(b)$ = there is a bus stop across the street from $b$
$C_3(b)$ = cars are approaching $b$
$C_4(b)$ = agent $b$ is about to cross the street
$A_1(b)$ = agent $b$ will cross the street
$A_2(a,b)$ = agent $a$ will pull $b$ off the street

*Cars are coming*

Agent $a$'s plan: $\Big(C_3(b) \wedge C_4(b)\Big) \Rightarrow_a A_2(a,b)$

Agent $b$'s plan: $\Big(C_1(b) \wedge C_2(b) \wedge \neg C_3(b)\Big) \Rightarrow_b A_1(b)$

There is no interference:

$$\Box_a \neg P\Big(A_1(b) \wedge A_2(a,b)\Big) \wedge \Diamond_a P\Big(C_1(b) \wedge C_2(b) \wedge C_3(b) \wedge \neg C_3(b) \wedge C_4(b)\Big)$$

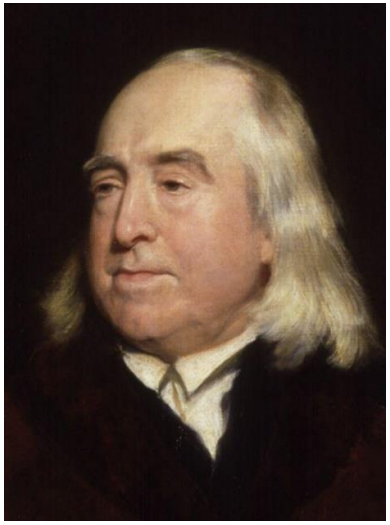
*False due to logical contradiction*

# Autonomy principle

- Why a strong "rationally constrained" provision?
  - *It is a consequence of the **deontological argument** for the autonomy principle.*
    - Strictly speaking, I adopt an **entire action policy** rather than individual action plans.
    - If I am to be rational, the plans must be **mutually consistent** (same for beliefs in general that I adopt).
    - Inconsistency is a strong condition: I am **rationally constrained** to acknowledge it.
    - The **universality of reason** says that when adopting a policy, I adopt it for **everyone** (Kant says I "legislate").
    - So, the action plans I rationally attribute to **everyone** must be mutually consistent.
    - If I adopt a plan that **conflicts** with the plans I rationally attribute to others, I am **rationally constrained** to acknowledge the inconsistency.
    - My policy is **irrational** and therefore **unethical**.

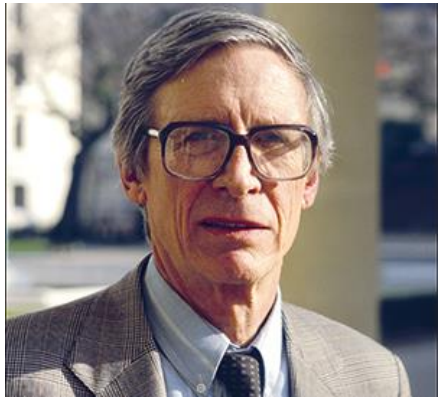
# Utilitarian principle

- This principle asks us to maximize total net expected "utility."
  - *As best we can estimate it.*
  - *"Greatest good for the greatest number," in Jeremy Bentham's formulation.*
  - *Utility = what the agent regards as* ***inherently valuable.***
    - That is, the end(s) to which one's actions are a means.
    - It was happiness/pleasure for classical utilitarians.
    - There must be an **ultimate end** to avoid infinite regress in the rationale for an act.



# Utilitarian principle

- Deontological argument – in brief.
  - *Due to **universality of reason**, if I regard an end as intrinsically valuable, I must regard it as valuable for **anyone**.*
    - It shouldn't matter who I am.
  - *My actions should take everyone else's utility as seriously as my own.*
    - This may not imply strict maximization of net expected utility.
    - For example, it may require some degree of distributive justice, as in the difference principle of John Rawls.



# Utilitarian principle

- What about **futility arguments**?
  - *My commanding officer orders me to torture a prisoner.*
    - The results are the same (or worse) if I refuse, as **someone else** will obey the order.
    - This shows that the torture passes the **utilitarian** test.

Abu Ghraib Prison, Iraq

# Utilitarian principle

- What about **futility arguments**?
  - *My commanding officer orders me to torture a prisoner.*
    - The results are the same (or worse) if I refuse, as **someone else** will obey the order.
    - This shows that the torture passes the **utilitarian** test.
  - *Yet it violates the prisoner's **autonomy**.*
    - The willingness of others to do it is irrelevant.
    - What matters is the **incompatibility** of action plans.

Abu Ghraib Prison, Iraq

# Logical formulation

Let social welfare function $W\big(C(a), A(a)\big)$ evaluate the expected utility distribution resulting from action plan $C(a) \Rightarrow_a A(a)$, which satisfies the utilitarian principle if and only if

$$\Diamond_a \forall A' \Big( W\big(C(a), A(a)\big) \geq W\big(C(a), A'(a)\big) \Big)$$

where $A'$ ranges over all otherwise ethical actions available to agent $a$ in circumstances $C(a)$.

*We move into 2nd order logic by quantifying over action predicates, but this can be avoided by introducing typed variables for actions.*

# Value alignment

- This is the incorporation of human values into AI-based decision making.
  - *But "values" is ambiguous.*
    - Values = what humans prefer
    - Values = what is preferable
  - *Value alignment normally uses **machine learning** to identify human **preferences**.*

# Value alignment

- This is the incorporation of human values into AI-based decision making.
  - *But "values" is ambiguous.*
    - Values = what humans prefer
    - Values = what is preferable
  - *Value alignment normally uses **machine learning** to identify human **preferences**.*
    - Example: MIT's "Moral Machine" learns preferred driving behavior by presenting scenarios to drivers worldwide.


  - *Our goal is to incorporate **ethics** as well: what is preferable.*

# Value alignment

- Goal: avoid **naturalistic fallacy** by combining empirical VA with independently derived ethical principles.
  - *Naturalistic fallacy = inferring "ought" from "is".*
    - For example, the fact that people prefer something doesn't imply they **should** prefer it.

David Hume

G. E. Moore

# Value alignment

- To evaluate an action plan in an AI rule base:
  - *Makes sure the antecedent is stated in full generality.*
  - *Apply the 3 ethical principles to the plan to generate 3* ***test propositions****.*
    - Each test proposition is a necessary condition for the plan to be ethical.
  - *Empirically determine the truth of the test propositions.*
    - By means of machine learning, etc.
  - *The action plan is ethical only if all 3 test propositions are* ***true****.*

# Value alignment

- Example.

$C_1(a) =$ An ambulance under the control of agent $a$ can reach its destination sooner by using the siren

$C_2(a) =$ There is an emergency patient in the ambulance.

$A(a) =$ The ambulance will use the siren.

*Consider the action plan:* $C_1(a) \Rightarrow_a A(a)$

*The generalization principle is*

$$\Diamond_a P\Big(\forall x\big(C(x) \to A(x)\big) \wedge C(a) \wedge A(a)\Big)$$

*This generates the test proposition*

$$\Diamond_a P\Big(\forall x\big(C_1(x) \to A(x)\big) \wedge C_1(a) \wedge A(a)\Big)$$

# Value alignment

- Example.

$C_1(a) =$ An ambulance under the control of agent $a$ can reach its destination sooner by using the siren
$C_2(a) =$ There is an emergency patient in the ambulance.
$A(a) =$ The ambulance will use the siren.

*Consider the action plan:* $C_1(a) \Rightarrow_a A(a)$

*The generalization principle is*

$$\Diamond_a P\Big(\forall x\big(C(x) \rightarrow A(x)\big) \wedge C(a) \wedge A(a)\Big)$$

*This generates the test proposition*

$$\Diamond_a P\Big(\forall x\big(C_1(x) \rightarrow A(x)\big) \wedge C_1(a) \wedge A(a)\Big)$$

*This is empirically **false**, since the agent cannot rationally believe that such general use of sirens would permit an ambulance to arrive sooner with a siren. **Violation.***

# Value alignment

- Example.

$C_1(a) =$ An ambulance under the control of agent $a$ can reach its destination sooner by using the siren

$C_2(a) =$ There is an emergency patient in the ambulance.

$A(a) =$ The ambulance will use the siren.

*Consider the action plan:* $\big(C_1(a) \wedge C_2(a)\big) \Rightarrow_a A(a)$

*The generalization principle is*

$$\Diamond_a P\Big(\forall x\big(C(x) \rightarrow A(x)\big) \wedge C(a) \wedge A(a)\Big)$$

*This generates the test proposition*

$$\Diamond_a P\Big(\forall x\big((C_1(x) \wedge C_2(x)) \rightarrow A(x)\big) \wedge C_1(a) \wedge C_2(a) \wedge A(a)\Big)$$

*This is empirically **true**, since evidence shows that ambulances can arrive sooner with a siren when it is always used for emergency transport.* ***No violation.***

# Value alignment

- Example that combines preferences with ethics.

$C_1(a) =$ Driver $a$ wishes to enter a main thoroughfare.
$C_2(a) =$ There are no gaps in the stream of traffic.
$A_1(a) =$ Driver $a$ will enter the main thoroughfare now.
$A_2(a) =$ Driver $a$ will wait for a gap in the traffic.

*Consider the action plan:* $\big(C_1(a) \wedge C_2(a)\big) \Rightarrow_a A_1(a)$

*The **utilitarian principle** generates the test proposition*

$$\Diamond_a \Big( W\big(C_1(a), C_2(a), A_1(a)\big) \geq W\big(C_1(a), C_2(a), A_2(a)\big) \Big)$$

# Value alignment

- Example that combines preferences with ethics.

$C_1(a) =$ Driver $a$ wishes to enter a main thoroughfare.
$C_2(a) =$ There are no gaps in the stream of traffic.
$A_1(a) =$ Driver $a$ will enter the main thoroughfare now.
$A_2(a) =$ Driver $a$ will wait for a gap in the traffic.

*Consider the action plan:* $\big(C_1(a) \wedge C_2(a)\big) \Rightarrow_a A_1(a)$

*The **utilitarian principle** generates the test proposition*

$$\Diamond_a \Big( W\big(C_1(a), C_2(a), A_1(a)\big) \geq W\big(C_1(a), C_2(a), A_2(a)\big) \Big)$$

*This is **false** in some Western countries, where drivers expect one to wait for a gap. Pulling into traffic risks an accident.*

# Value alignment

- Example that combines preferences with ethics.

$C_1(a) = \text{Driver } a \text{ wishes to enter a main thoroughfare.}$
$C_2(a) = \text{There are no gaps in the stream of traffic.}$
$A_1(a) = \text{Driver } a \text{ will enter the main thoroughfare now.}$
$A_2(a) = \text{Driver } a \text{ will wait for a gap in the traffic.}$

*Consider the action plan:* $\big(C_1(a) \wedge C_2(a)\big) \Rightarrow_a A_1(a)$

*The **utilitarian principle** generates the test proposition*

$$\Diamond_a \Big( W\big(C_1(a), C_2(a), A_1(a)\big) \geq W\big(C_1(a), C_2(a), A_2(a)\big) \Big)$$

*This is **false** in some Western countries, where drivers expect one to wait for a gap. Pulling into traffic risks an accident.*

*It may be **true** in some other areas, where drivers make allowances for entering traffic.*

***Empirical value alignment (ML)** can resolve the issue.*

# Value alignment

- Example involving a nursing home robot.
  - Similar to an example in Anderson and Anderson (2015).
  - *A robot dispenses medications to a nursing home patient.*
    - The patient **refuses** to take the pills.
    - The robot is programmed to **report** this to the head nurse.
    - This will result in **confinement** to a certain ward, because the pills prevent dangerous disorientation.

# Value alignment

- Example involving a nursing home robot.
  - Similar to an example in Anderson and Anderson (2015).
  - *A robot dispenses medications to a nursing home patient.*
    - The patient **refuses** to take the pills.
    - The robot is programmed to **report** this to the head nurse.
    - This will result in **confinement** to a certain ward, because the pills prevent dangerous disorientation.
  - *The patient complains that the nursing home **violates her autonomy**, because she wants to visit a relative.*
    - Autonomy principle doesn't require us to allow people to do **whatever they want**.
    - However, confinement to a ward is **coercion**.
    - On entering the nursing home, the patient signed a **consent form** with full awareness and understanding of nursing home policy.

# Value alignment

$C_1(b) =$ Patient $b$ takes the pills.
$C_2(b) =$ Patient $b$ signed the consent form.
$C_3(b) =$ Patient $b$ wants to visit relatives.
$A_1(a) =$ Robot $a$ informs the head nurse.
$A_2(b) =$ Patient $b$ visits relatives.

*The robot's action plan:* $\big(\neg C_1(b) \wedge C_2(b)\big) \Rightarrow_a A_1(a)$

*The patient's action plan:* $\Big(\big(C_1(b) \vee \neg C_2(b)\big) \wedge C_3(b)\Big) \Rightarrow_b A_2(b)$

*We have interference if*

$$\boxed{\Box_a \neg P\big(A_1(a) \wedge A_2(b)\big) \wedge}$$
$$\Diamond P\Big(\neg C_1(b) \wedge C_2(b) \wedge \big(C_1(b) \vee \neg C_2(b)\big) \wedge C_3(b)\Big)$$

***True*** *because nursing home prohibits excursions when patient refuses the pills*

# Value alignment

$C_1(b) =$ Patient $b$ takes the pills.
$C_2(b) =$ Patient $b$ signed the consent form.
$C_3(b) =$ Patient $b$ wants to visit relatives.
$A_1(a) =$ Robot $a$ informs the head nurse.
$A_2(b) =$ Patient $b$ visits relatives.

*The robot's action plan:* $\big(\neg C_1(b) \wedge C_2(b)\big) \Rightarrow_a A_1(a)$

*The patient's action plan:* $\Big(\big(C_1(b) \vee \neg C_2(b)\big) \wedge C_3(b)\Big) \Rightarrow_b A_2(b)$

*We have interference if*

$$\Box_a \neg P\big(A_1(a) \wedge A_2(b)\big) \wedge$$
$$\Diamond_a P\Big(\neg C_1(b) \wedge C_2(b) \wedge \big(C_1(b) \vee \neg C_2(b)\big) \wedge C_3(b)\Big)$$

***False*** *because one cannot rationally believe a logical contradiction*

# Value alignment

$C_1(b) =$ Patient $b$ takes the pills.
$C_2(b) =$ Patient $b$ signed the consent form.
$C_3(b) =$ Patient $b$ wants to visit relatives.
$A_1(a) =$ Robot $a$ informs the head nurse.
$A_2(b) =$ Patient $b$ visits relatives.

*The robot's action plan:* $\big(\neg C_1(b) \wedge C_2(b)\big) \Rightarrow_a A_1(a)$

*The patient's action plan:* $\Big(\big(C_1(b) \vee \neg C_2(b)\big) \wedge C_3(b)\Big) \Rightarrow_b A_2(b)$

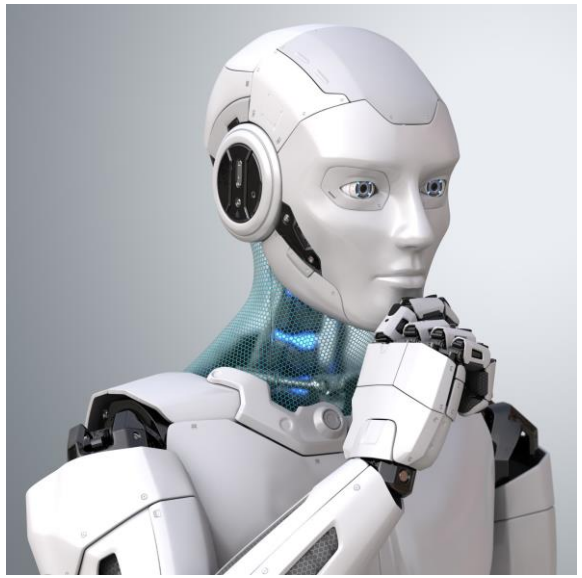
*We have interference if*

$$\Box_a \neg P\big(A_1(a) \wedge A_2(b)\big) \wedge$$
$$\Diamond_a P\Big(\neg C_1(b) \wedge C_2(b) \wedge \big(C_1(b) \vee \neg C_2(b)\big) \wedge C_3(b)\Big)$$

*So there is* ***no autonomy violation****.*

# Postscript

- Nothing in deontological ethics presupposes that agents are **human**.
  - *A reasons-responsive machine can, in principle, be an **autonomous agent**.*
    - It **explains** the rationale for its actions on demand.
    - It doesn't matter if its actions are determined by a **program** (**our** actions are determined).
  - *It can have **obligations** to us, and we to it.*
    - Although **utilitarian** obligations are tricky for machines.
    - Since they are **nonhuman**.



# References

J. N. Hooker and T. W. Kim, Toward non-intuition-based machine and AI ethics: A deontological approach based on modal logic, *AIES 2018*, 130-136.

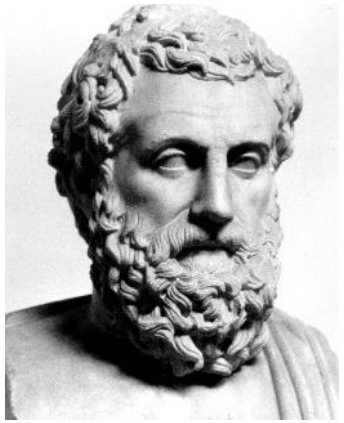
J. N. Hooker and T. W. Kim, Truly autonomous machines are ethical, *AI Magazine* **40** (2020) 66-73.

T. W. Kim, J. N. Hooker, and T. Donaldson, Taking principles seriously: A hybrid approach to value alignment in artificial intelligence, *Journal of AI Research* **70** (2021) 871-890.

# Conscious rationale?

- A flaw in rationality-based ethics?
  - *Most of our actions are not consciously justified.*
    - We can't devise a rationale for everything we do.
    - We are creatures of habit.
  - ***Dual process theory** agrees.*
    - **System 1 thinking** is fast and unconscious.
    - **System 2 thinking** is slow and based on conscious reasoning.
    - We usually rely on System 1.
    - *Kahneman (2011)*


# Conscious rationale?



- Ethicists are well aware of this
  - *Going back at least to Aristotle.*
  - *We deliberately **initiate** habits.*
  - *We allows habits to **continue**.*
    - If I continue smoking, I **decide** not to break the habit.
  - *We can **invoke system 2 thinking** when needed.*
    - Part of being ethical is being **autonomous agents**.
    - That is, making conscious decisions based on reasons at strategic junctures.