Reasoning about Moral Conflicts in AI

MARIJA SLAVKOVIK
UNIVERSITY OF BERGEN
MARIJA.SLAVKOVIK@UIB.NO

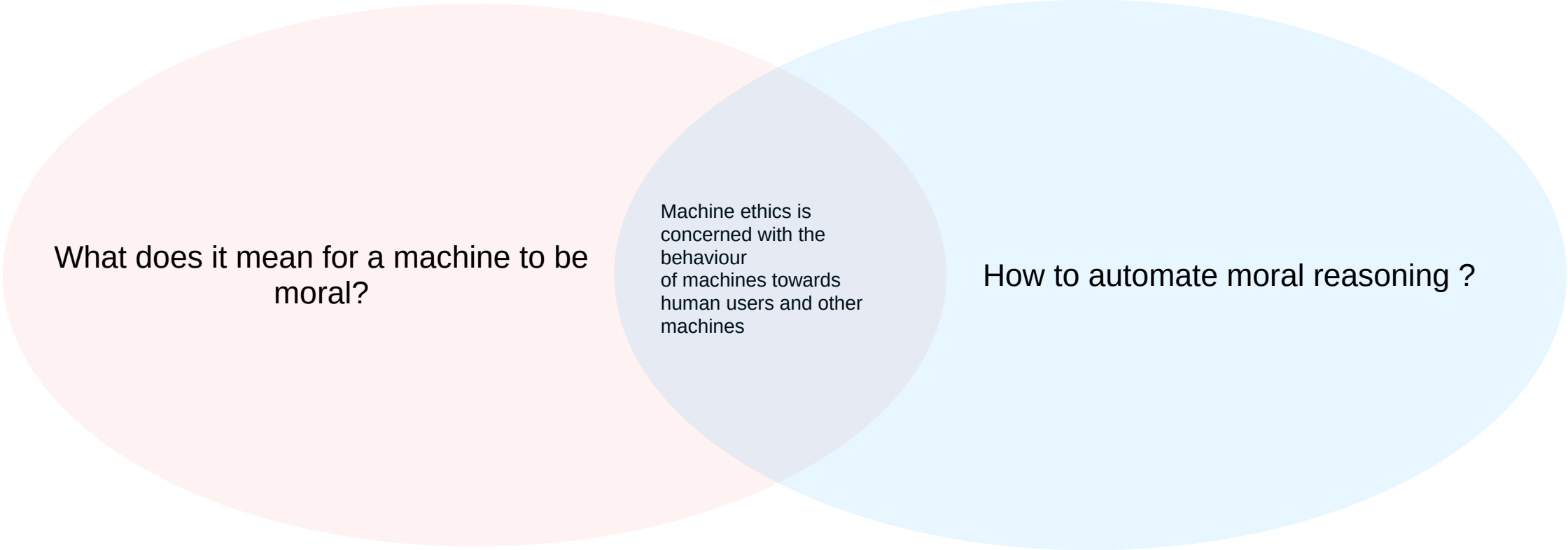
AI ethics

How to ensure no negative ethical footprint of AI in society?

Machine ethics

“is concerned with the behaviour of machines towards human users and other machines”

Machine ethics



Machines as moral arbiters

The decision making process

Decision making is a process than consists of:

1. identify the problem for which a decision needs to be made,
2. evaluate the objectives and preferences that apply,
3.  analyse the decision problem and its constraints, and develop or identify the possible options from which to choose,
4.  choose from the identified options following some reasoning.

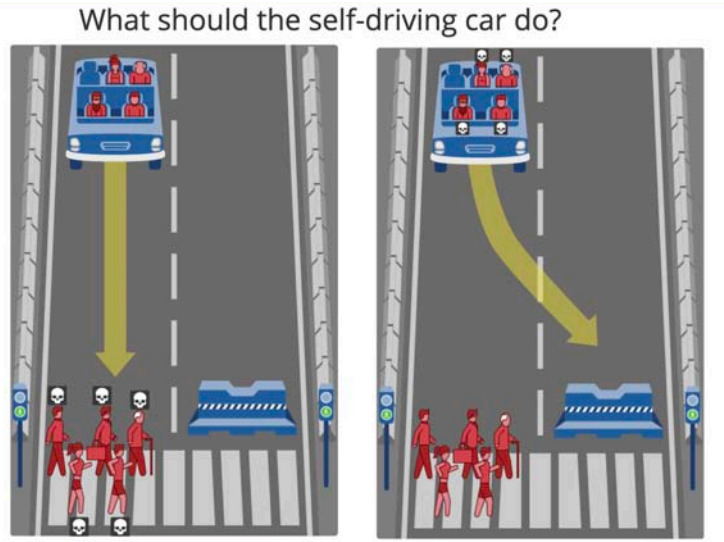
"The greater the freedom of a machine, the more it will need moral standards." Picard R (1997) Affective computing. MIT Press, Cambridge

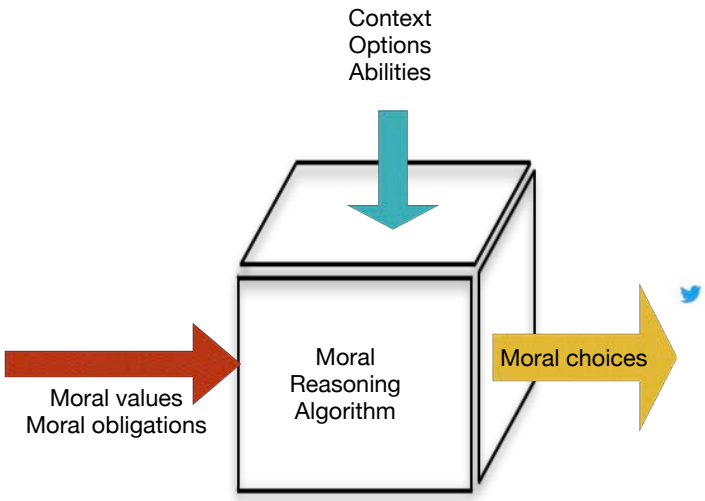
Moral decisions

- A **moral decision** is a choice made based not only on the factual objectives, preferences and constraints, but also based on a person's or societie's consideration of what is moral behaviour.
- Moral decisions also include considering "the interests of others as of equal weight with one's own"

Machines ethics

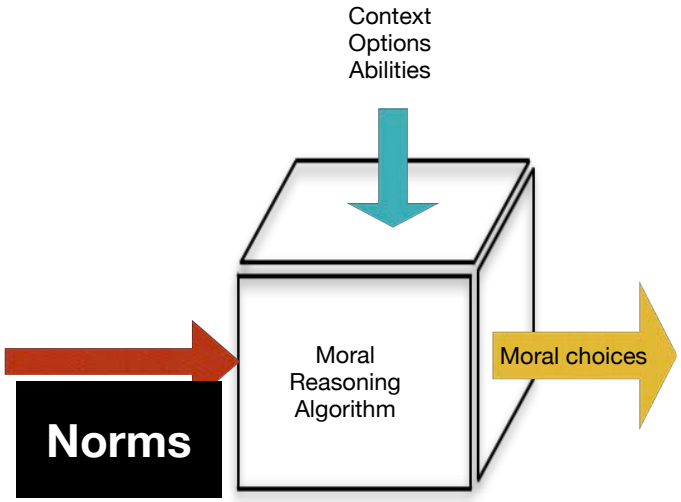
How to reason in a morally sensitive context?





Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
	Propensity for Harm	

But.. isn't this normative reasoning?



Published: March 1999

Introduction: Agents and Norms: How to fill the gap?

Rosaria Conte, Rino Falcone & Giovanni Sartor

Artificial Intelligence and Law 7, 1–15(1999) | Cite this article

Normative (multi-)agent systems:

- Norm-governed interaction of autonomous systems
- How agents can acquire norms?
- How agents can violate norms?
- How an agent can be autonomous?

Normative reasoning and machine ethics

The same but different

TABLE 1

A TYPOLOGY OF NORMS

		Low probability that an attempt will be made to apply a sanction* when the act occurs†	High probability that an attempt will be made to apply a sanction* when the act occurs†			
			By anyone (i.e., without regard to status)		Only by a person or persons in a particular status or statuses	
			By means that exclude the use of force	By means that may include the use of force	By means that exclude the use of force	By means that may include the use of force
Collective evaluation of the act‡	Collective expectation concerning the act§	Type A: Collective conventions	Type D: Collective morals	Type H: Collective mores	Type L: Collective rules	Type P: Collective laws
	No collective expectation concerning the act	Type B: Problematic conventions	Type E: Problematic morals	Type I: Problematic mores	Type M: Problematic rules	Type Q: Problematic laws
No collective evaluation of the act	Collective expectation concerning the act§	Type C: Customs	Type F: Possible empirical null class	Type J: Possible empirical null class	Type N: Exogenous rules	Type R: Exogenous laws
	No collective expectation concerning the act	Logical null class, i.e., non-normative	Type G: Possible empirical null class	Type K: Possible empirical null class	Type O: Coercive rules	Type S: Coercive laws

Norms: The Problem of Definition and Classification

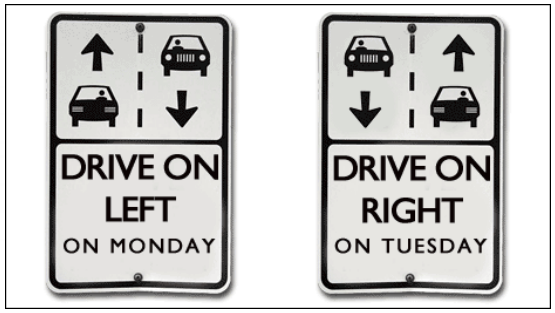
Author(s): Jack P. Gibbs

Source: *American Journal of Sociology*, Mar., 1965, Vol. 70, No. 5 (Mar., 1965), pp. 586-594

Published by: The University of Chicago Press

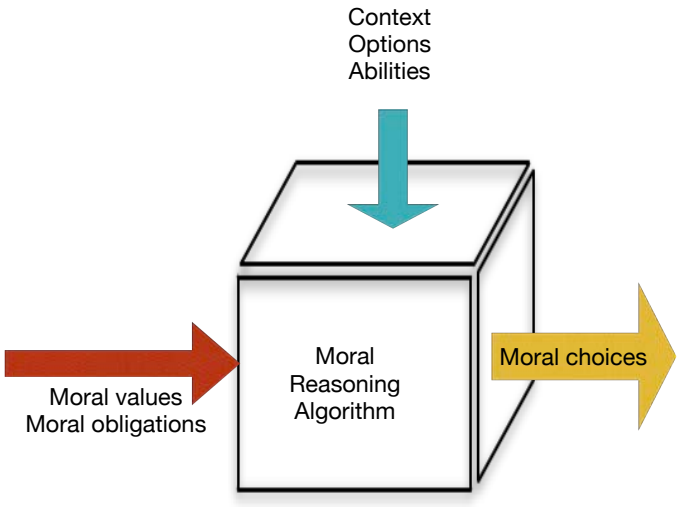
Stable URL: https://www.jstor.org/stable/2774978

- Following norms are not the only way to achieve moral behaviour

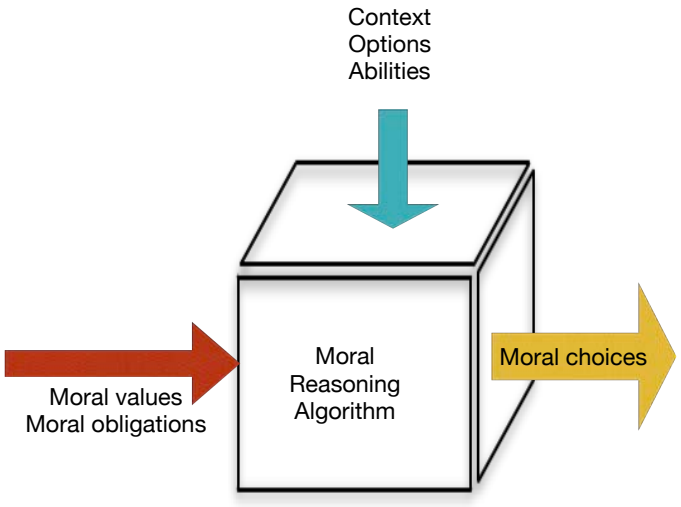


- Not all norms are moral

How do we do it?

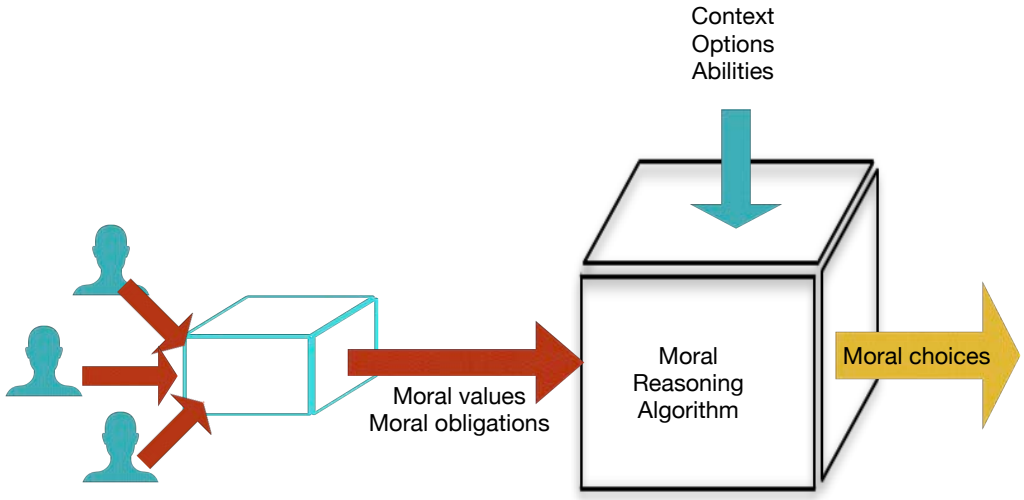


What can we do?



Machines ethics

Who supplies the moral information?



Moral disagreement

Moral Disagreement and Artificial Intelligence.

Pamela Robinson.

- The methodological problem: How should we design artificially intelligent systems that align with morality or our values when neither the designers nor those affected by these systems can agree about what's moral or valuable?

Moral conflicts

- To program a machine to do the right thing we need to know what the right thing is

For one thing, the task of actually applying a correct moral theory to each of the ethical decisions we face every day would be difficult and time-consuming; and it seems unlikely, for most of us, that such a theory could have any more bearing upon our day to day ethical reasoning than physics has upon our everyday reasoning about objects in the world. Most of our common-sense ethical thinking seems to be guided instead, not by the dictates of moral theory, but by simple rules of thumb – 'Return what you borrow', Don't cause harm' – and it is not hard to generate conflicts among these.[3]

Published: February 1994

Moral dilemmas and nonmonotonic logic

John F. Horty

Journal of Philosophical Logic **23**, 35–65(1994) | Cite this article

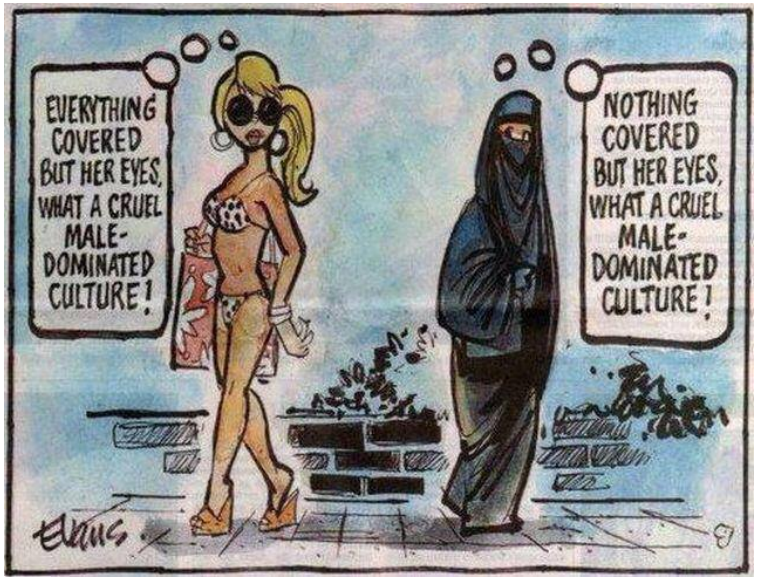
Ideal advisors vs whose life is it anyways

- Tech colonialism vs ethical relativisam

Economics and Philosophy, **32** (2016) 283–321 © Cambridge University Press
doi:10.1017/S0266267115000486 First published online 11 January 2016
journals.cambridge.org/eap

AGGREGATING MORAL PREFERENCES

MATTHEW D. ADLER*



OK, so it is a collective decision

Implementations of social choice ethics must make three types of choices, each of which create their own set of ethical dilemmas (Baum 2009):

1. *Standing* Who or what is included in the group to have its values factored into the AI?
2. *Measurement* What procedure is used to obtain values from each member of the selected group?
3. *Aggregation* How are the values of individual group members combined to form the aggregated group values?

AI & Soc (2020) 35:165–176
DOI 10.1007/s00146-017-0760-1

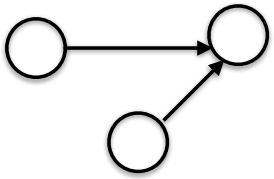
ORIGINAL ARTICLE

Social choice ethics in artificial intelligence

Seth D. Baum[1]

..but there is more

$\forall P.[P(x,y) \leftrightarrow P(y,x)]$



Example	Input Attributes										Goal
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

Figure 18.3 Examples for the restaurant domain.

1. What we elicit influences what conflicts can arise.
2. Moral views vs moral obligations vs moral values vs moral theories: each of these has a different KR formalism.
3. KR formalism influences agreement/aggregation/resolution algorithm choice.

arxiv.org/abs/1812.04741

Cornell University

arXiv.org > cs > arXiv:1812.04741

Search... Help | Advanc

Computer Science > Artificial Intelligence

[Submitted on 11 Dec 2018 (v1), last revised 6 Mar 2019 (this version, v2)]

Building Jiminy Cricket: An Architecture for Moral Agreements Among Stakeholders

Beishui Liao, Marija Slavkovik, Leendert van der Torre

An autonomous system is constructed by a manufacturer, operates in a society subject to norms and laws, and is interacting with end-users. We address the challenge of how the moral values and views of all stakeholders can be integrated and reflected in the moral behaviour of the autonomous system. We propose an artificial moral agent architecture that uses techniques from normative systems and formal argumentation to reach moral agreements among stakeholders. We show how our architecture can be used not only for ethical practical reasoning and collaborative decision-making, but also for the explanation of such moral behavior.

Comments: Presented at the AAAI/ACM Artificial Intelligence, Ethics and Society
Subjects: **Artificial Intelligence (cs.AI)**
Cite as: arXiv:1812.04741 **[cs.AI]**
(or arXiv:1812.04741v2 **[cs.AI]** for this version)

Submission history

From: Marija Slavkovik [view email]
[v1] Tue, 11 Dec 2018 23:16:16 UTC (1,286 KB)
[v2] Wed, 6 Mar 2019 15:23:15 UTC (1,286 KB)

Argumentation

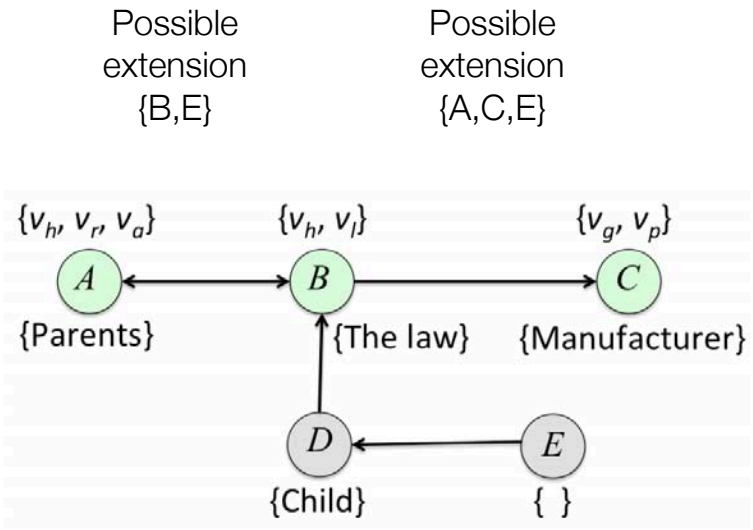
Normative reasoning

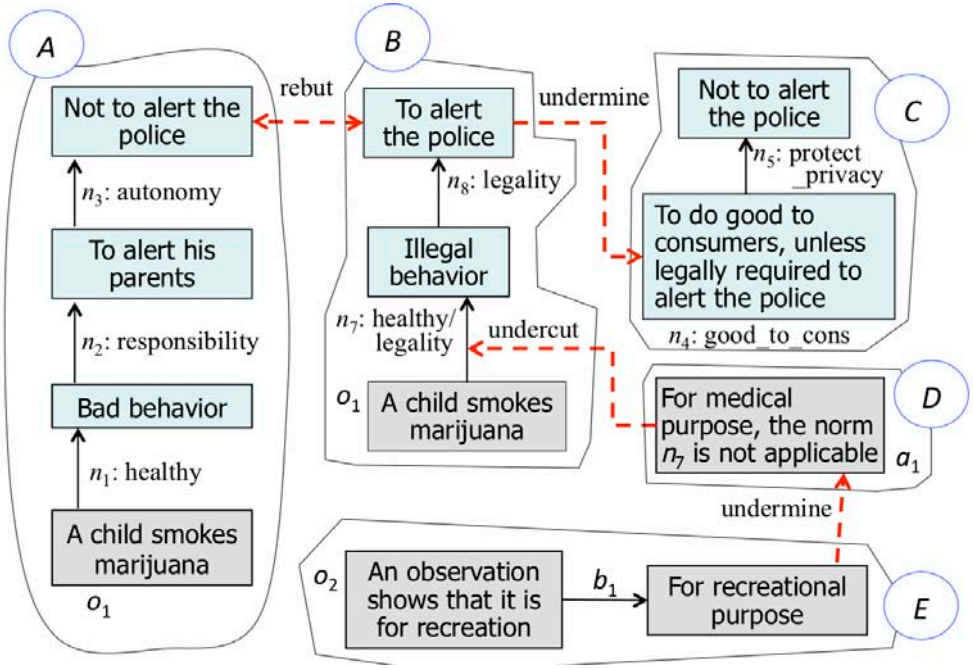
The idea

1. We ask stakeholders what they value/what duties they want to respect before machine is deployed
2. Machine uses stakeholder values to build arguments in running time
3. Machine simulates an argumentation whenever there is decision to be made
4. Machine uses argumentation theory to find out what to do

How do we build arguments?

- Each stakeholder is represented with a set of values

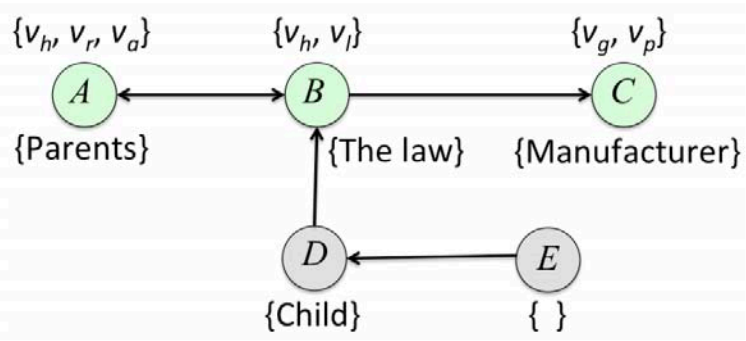




- How do we know which extension to choose?

How do we resolve

- Since values are degrees of importance of some things or actions, one may argue that a reasonable solution is to accept the extension that reaches the maximal extent of agreement over a set of values.
- For an extension E ⊆ A associated with a set of value V_E we say that it reaches the maximal extent of agreement over V iff there is no another extension E' ⊆ A associated with a set of values $V_{E'}$ s.t. $V_{E'}$ has a higher priority over V_E , denoted as $V_{E'} > V_E$.



Possible extension {B,E}

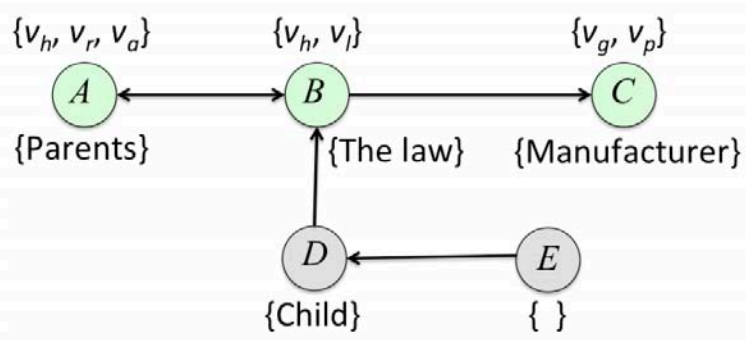
$\{V_h, V_l\}$

Possible extension {A,C,E}

$\{V_h, V_r, V_a, V_g, V_p\}$

How do we resolve

- Since values are degrees of importance of some things or actions, one may argue that a reasonable solution is to accept the extension that reaches the maximal extent of agreement over a set of values.
- For an extension E ⊆ A associated with a set of value V_E we say that it reaches the maximal extent of agreement over V iff there is no another extension E' ⊆ A associated with a set of values $V_{E'}$ s.t. $V_{E'}$ has a higher priority over V_E , denoted as $V_{E'} > V_E$.



Possible extension {B,E}

$\{V_h, V_l\}$

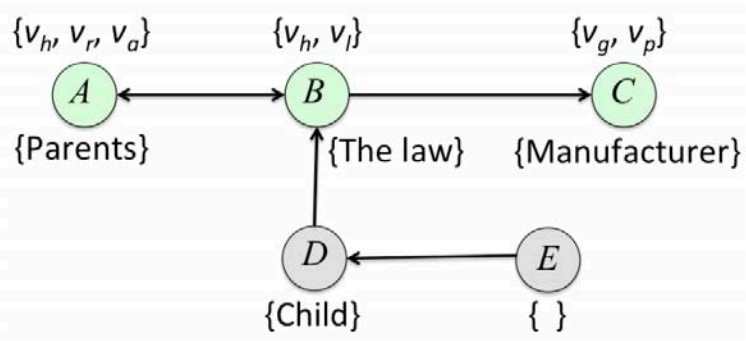
$>$

Possible extension {A,C,E}

$\{V_h, V_r, V_a, V_g, V_p\}$

How do we resolve

- Since values are degrees of importance of some things or actions, one may argue that a reasonable solution is to accept the extension that reaches the maximal extent of agreement over a set of values.
- For an extension E ⊆ A associated with a set of value V_E we say that it reaches the maximal extent of agreement over V iff there is no another extension E' ⊆ A associated with a set of values $V_{E'}$ s.t. $V_{E'}$ has a higher priority over V_E , denoted as $V_{E'} > V_E$.



Possible extension {B,E}

$\{V_h, V_l\}$

<

Possible extension {A,C,E}

$\{V_h, V_r, V_a, V_g, V_p\}$

How do we resolve conflicts?

- The priority relation between two sets of values can be defined in term of a partial ordering over V and a lifting principle, e.g., **the elitist principle** or **the democratic principle** Modgil and Prakken (2013)
- Assume we are given a partial ordering over V by using $v_1 \geq v_2$ to denote v_1 is at least as good as v_2, and two sets $V_1 \subseteq V$ and $V_2 \subseteq V$.
- The elitist principle can be defined as: $V_1 \geq V_2$ iff there exists $v \in V_2$ such that $v' \geq v$ for all $v' \in V_1$.
- The democratic principle can be defined as: $V_1 \geq V_2$ iff for all $v \in V_2$ there exists $v' \in V_1$ such that $v' \geq v$.

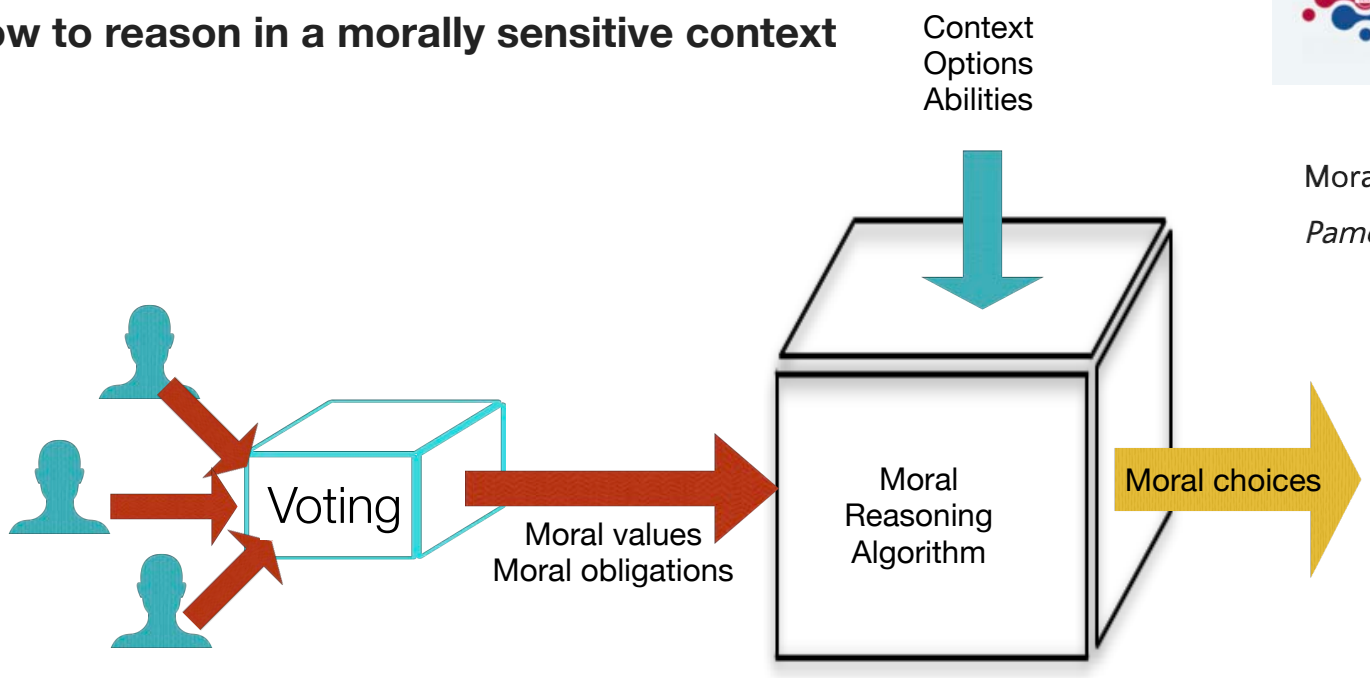
How do we resolve conflicts?

- The priority relation between two sets of values can be defined in term of a partial ordering over V and a lifting principle, e.g., **the elitist principle** or **the democratic principle** Modgil and Prakken (2013)
- Assume we are given a partial ordering over V by using $v_1 \geq v_2$ to denote v_1 is at least as good as v_2, and two sets $V_1 \subseteq V$ and $V_2 \subseteq V$.
- The elitist principle can be defined as: $V_1 \geq V_2$ iff there exists $v \in V_2$ such that $v' \geq v$ for all $v' \in V_1$.
- The democratic principle can be defined as: $V_1 \geq V_2$ iff for all $v \in V_2$ there exists $v' \in V_1$ such that $v' \geq v$.

Moral philosophy

Machines ethics

How to reason in a morally sensitive context



AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE,
ETHICS, AND SOCIETY

Moral Disagreement and Artificial Intelligence.

Pamela Robinson.

[…] what makes moral disagreement especially challenging is that there are two very different ways of handling it. **Political** solutions aim for a fair compromise, while **epistemic** solutions aim at moral truth.

Majority aggregation

- As many as possible should get what they want
- It only works if everyone has a chance to become part of the majority.
- How often is aggregation on moral views to happen? Once? Every 4 years?
- How small should a minority be for its moral views to be irrelevant for the aggregation?

~~What should voting be like?~~

~~What should judgment aggregation be like?~~

Egalitarian Judgment Aggregation

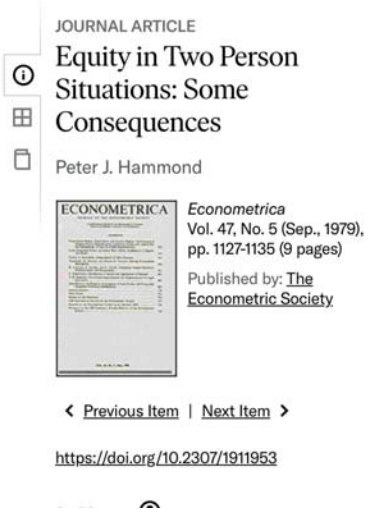
Sirin Botan, Ronald de Haan, Marija Slavkovik and Zoi Terzopoulou

Maximin property

A rule F satisfies the **maximin** property if for all profiles $\mathbf{J} \in \mathcal{J}(\Phi)^n$ and judgments $J \in F(\mathbf{J})$ there do not exist judgment $J' \in \mathcal{J}(\Phi)$ and agent $j \in N$ such that

$$H(J_i, J') < H(J_j, J) \text{ for all } i \in N.$$

If person *i* is worse off than person *j* both in outcome **x** and in outcome **y**, and if *i* is better off himself in x than in y, while *j* is better off in **y** than in **x**, and if furthermore all others are just as well off in **x** as in **y**, then **x** is socially at least as good as **y**.


JOURNAL ARTICLE

Equity in Two Person Situations: Some Consequences

Peter J. Hammond

ECONOMETRICA

Econometrica
Vol. 47, No. 5 (Sep., 1979), pp. 1127-1135 (9 pages)

Published by: The Econometric Society

‹ Previous Item | Next Item ›

https://doi.org/10.2307/1911953

Equity property

A rule F satisfies the **equity** property if for all profiles $\mathbf{J} \in \mathcal{J}^n$ and judgments $J \in F(\mathbf{J})$, there do not exist judgment $J' \in \mathcal{J}(\Phi)$ and agents $i', j' \in N$ such that

$$|H(J_i, J') - H(J_j, J')| < |H(J_{i'}, J) - H(J_{j'}, J)| \text{ for all } i, j \in N.$$

Inequalities are decreased when we transfer from the most satisfied agent to the least satisfied agent

Property relations

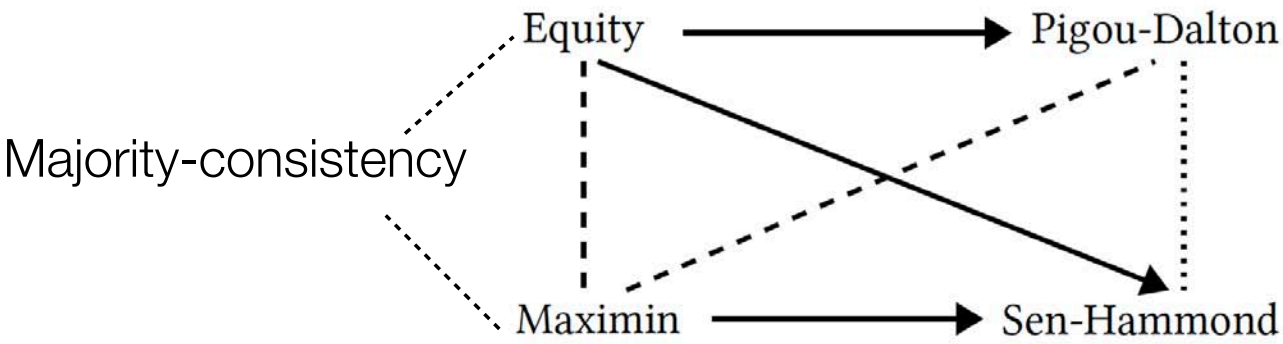
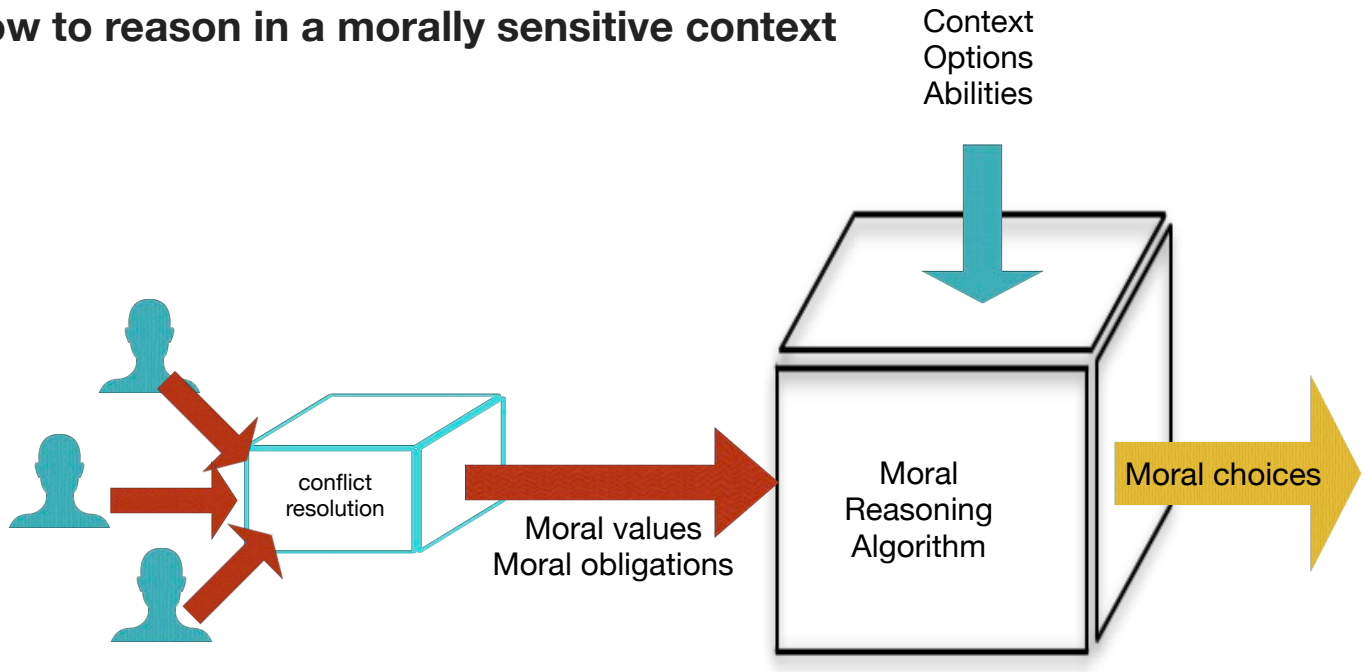


Figure 1: Dashed lines denote incompatibility, dotted lines incomparability, and arrows implication relations.

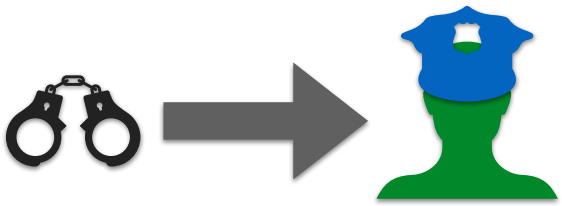
Machines ethics

How to reason in a morally sensitive context



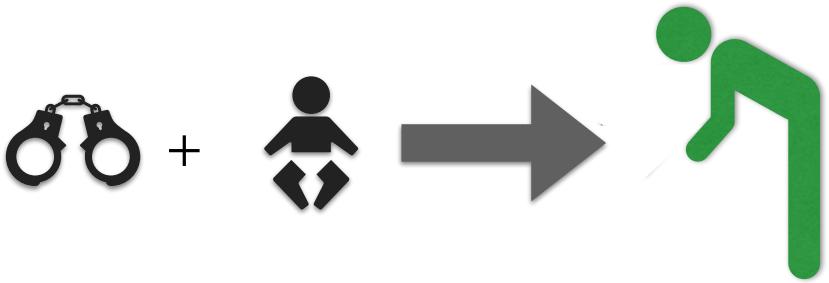
Can we find a compromise?

- Focus on norms: If x, one should do y.
- Focus on compromise: each of the stakeholders makes concessions to their moral view.
- Use the *lex specialis derogat legi generali* legal principle



Can we find a compromise?

- Focus on norms: If x, one should do y.
- Focus on compromise: each of the stakeholders makes concessions to their moral view.
- Use the *lex specialis derogat legi generali* legal principle

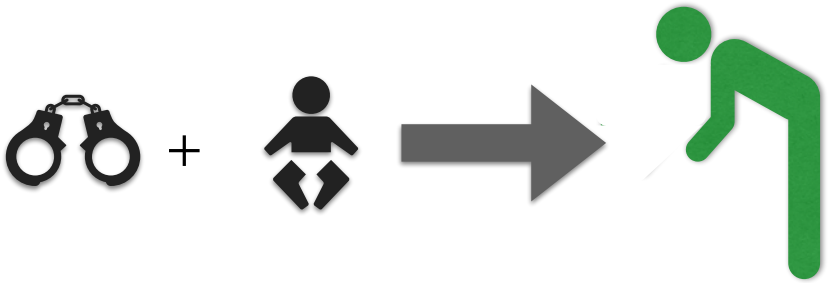


Can we find a compromise?

Ana Ozaki

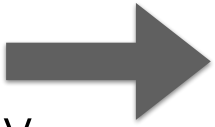
In Progress

- Focus on norms: If x, one should do y.
- Focus on compromise: each of the stakeholders makes concessions to their moral view.
- Use the *lex specialis derogat legi generali* legal principle



The algorithm

- If x, one should do y.
- If x and z, one should do ¬y.



- If x and not z, one should do y.
- If x and z, one should do ¬y.

- If x, one should do y. 
- If x, one should do ¬y.

Postulates defining what is a compromise

- P1: The compromise is coherent, no two norms advising “opposite” actions
- P2: If the union of the norms is coherent, then that is the compromise
- P3: No one’s norm is fully “overridden” by the compromise. An input “If x, then z” cannot become “If x then ¬z” in the compromise
- P4: Every norm in the compromise has an origin in a norm proposed by a stakeholder
- P5: Every norm from each stakeholder has a norm that “represents it” in the compromise
- P6: Norms are only “weakened”/“made more specific” by a “relevant” condition
- P7: The compromise is as “general” as possible

Where are we in this?

Thank you