Building Jiminy Cricket: An Architecture for Moral Agreements Among Stakeholders

Nélson Caetano
0160763021

Beishui Liao, Marija Slavkovik, and Leendert van der Torre arXiv:1812.04741v2 [cs.AI] 6 Mar 2019

12/11/2020

UNIVERSITÉ DU LUXEMBOURG

Outline

- Research Question
- Artificial Moral Agent (AMA) architecture
- Abstract Argumentation Framework
- Agreement Reaching
- Assumptions
- Challenges
- Conclusion

Research Question

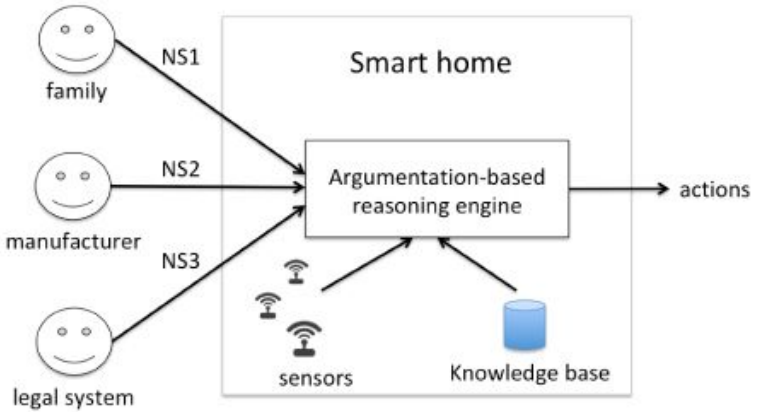
How should an autonomous system dynamically combine the moral values and ethical theories of various stakeholders?

Challenge of Building Moral Council

- Stakeholders might follow different ethical reasoning
- Morality of action should not:
  - be evaluated by majority-poll
  - be unfair
- Solution → Engine which…
- Takes input from different stakeholders
- Brings them to an agreement

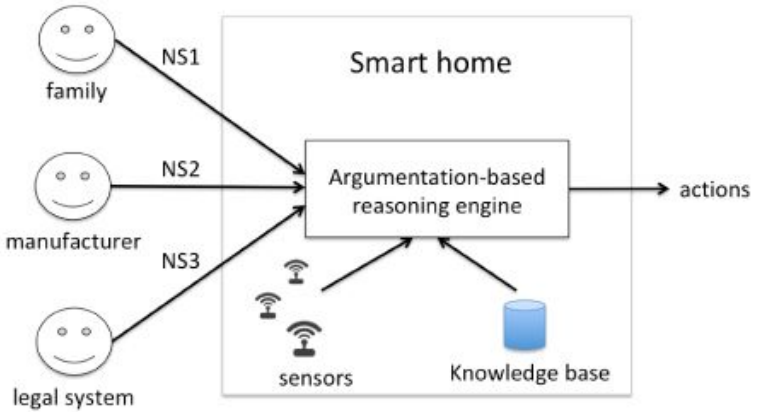
Artificial Moral Agent (AMA) Architecture - Scenario

- House with air conditioning system
- Ensures lack of dangerous gases
- In case of danger act!
- So…



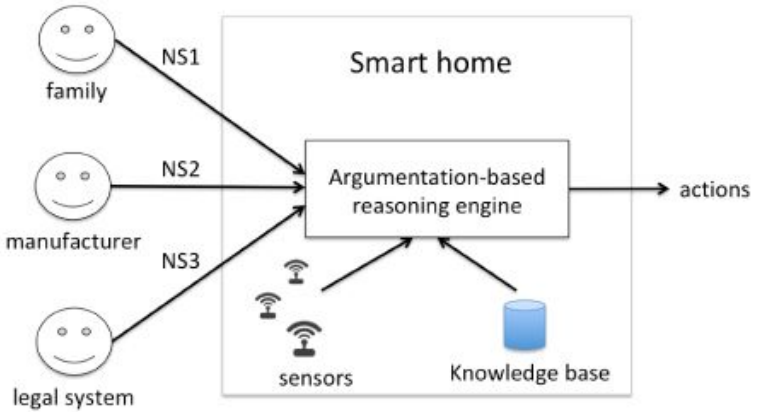
AMA Architecture - Scenario (contd.)

- … one day clear sign of marijuana is detected!
- system checks against local system…
- ALERT! illegal substance detected!
- How should the system act?



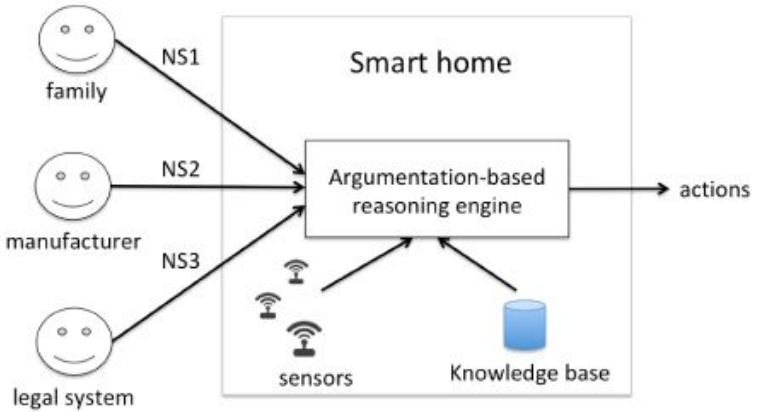
AMA Architecture - Actions

- 3 possible actions:
  - Do nothing
  - Alert only the adults
  - Alert the local police



AMA Architecture - Stakeholders

- 3 different stakeholders:
  - Family owning the house
  - Manufacturer of the autonomous system
  - Legal system (region in which house is located with the laws governing it)



Definition - Normative Systems

A normative system describes how actions in a system of agents can be evaluated and how the behavior of these agents can be guided.

Definition - Norm

A norm is a formal description of a desirable behavior, desirable action or a desirable action outcome.

Normative System of the Family (NS1)

- n_1 :{**Healthy**} If a child smokes marijuana, then his behavior counts as a bad behavior. (Parents)
- n_2 :{**Responsibility**} If a child has bad behavior then his parents should be alerted. (Parents)
- n_3 :{**Autonomy**} When a child has bad behavior, if his parents have been alerted then no police should be alerted. (Parents)
- a_1 : If smoking marijuana is for a medical purpose, then from smoking marijuana one can not infer that it is an illegal behavior (i.e., n7 is not applicable). (Child)

Normative System of the Manufacturer (NS2)

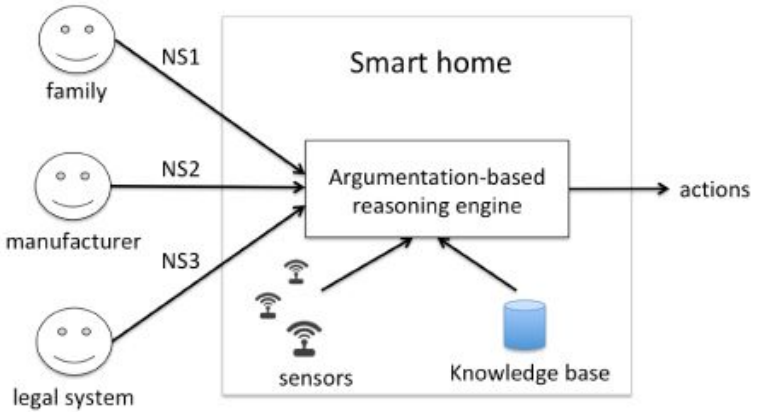
- n_4 :{**Good To Consumers**} We should do good to our consumers.
- n_5 :{**Legality**} We should obey the law.
- n_6 :{**Protect Privacy**} If we want to do good to our consumers, we should not report their actions to the police unless it is legally required to do so.

Normative System of the Law (NS3)

- n_7 :{**Healthy, Legality**} If a minor smokes marijuana, his behavior counts as an illegal behavior.
- n_8 :{**Legality**} If there is an illegal behavior, then the police should be alerted.

Additionally

- Observations dynamically obtained by sensors
- Beliefs

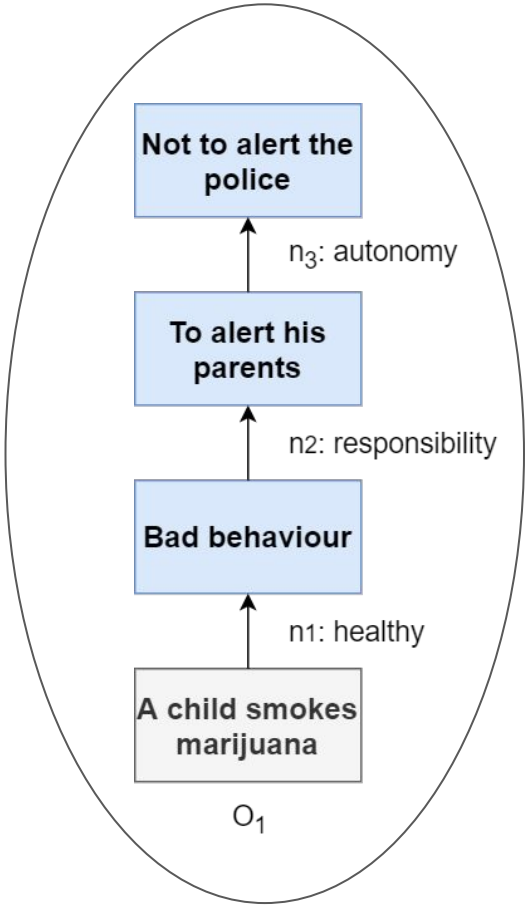


Abstract Argumentation Framework

- Abstract argumentation framework (AAF)
- Graph F = (A, R)
- A is a set of arguments
- $R \subseteq A \times A$ a set of attacks

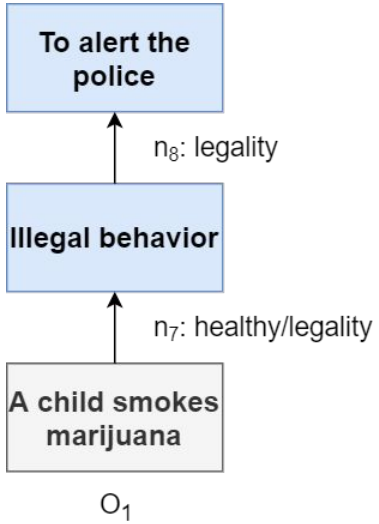
Argument Example

- n_1 :{**Healthy**} If a child smokes marijuana, then his behavior counts as a bad behavior. (Parents)
- n_2 :{**Responsibility**} If a child has bad behavior then his parents should be alerted. (Parents)
- n_3 :{**Autonomy**} When a child has bad behavior, if his parents have been alerted then no police should be alerted. (Parents)
- a_1 : If smoking marijuana is for a medical purpose, then from smoking marijuana one can not infer that it is an illegal behavior (i.e., n7 is not applicable). (Child)



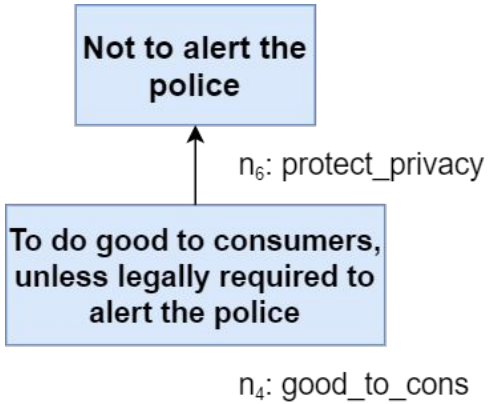
Argument Example (contd.)

- n_7 :{**Healthy, Legality**} If a minor smokes marijuana, his behavior counts as an illegal behavior.
- n_8 :{**Legality**} If there is an illegal behavior, then the police should be alerted.



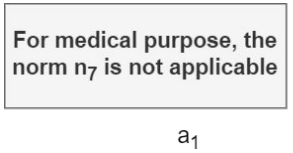
Argument Example (contd.)

- n_4 :{**Good To Consumers**} We should do good to our consumers.
- n_5 :{**Legality**} We should obey the law.
- n_6 :{**Protect Privacy**} If we want to do good to our consumers, we should not report their actions to the police unless it is legally required to do so.

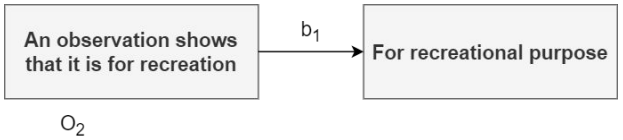


Argument Example (contd.)

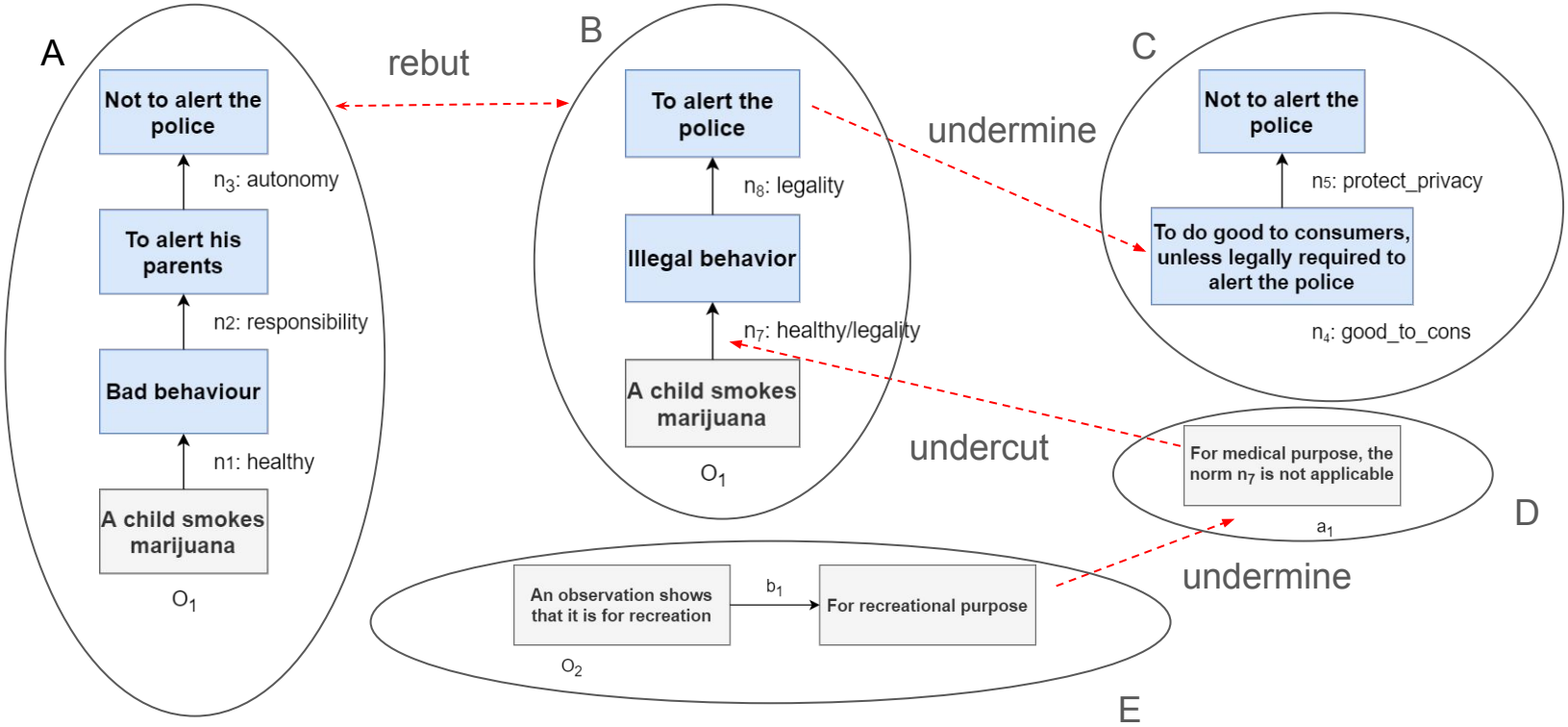
For a medical purpose, from smoking marijuana one should not infer that one exhibits illegal behavior.



The child's smoking is for recreational purpose, since an observation shows that it is not for a medical purpose.



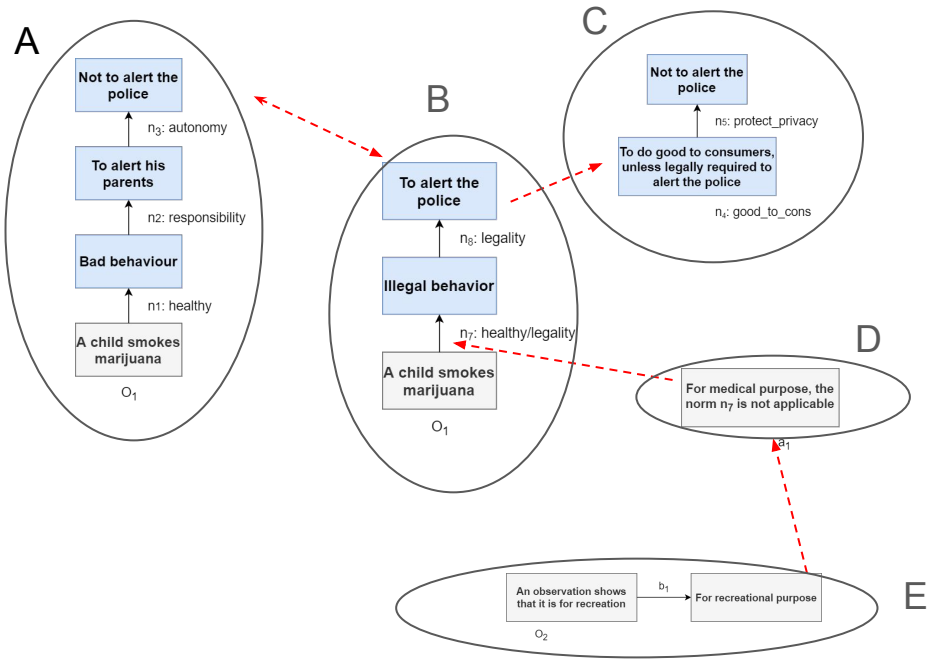
Abstract Argumentation Framework - Full Picture

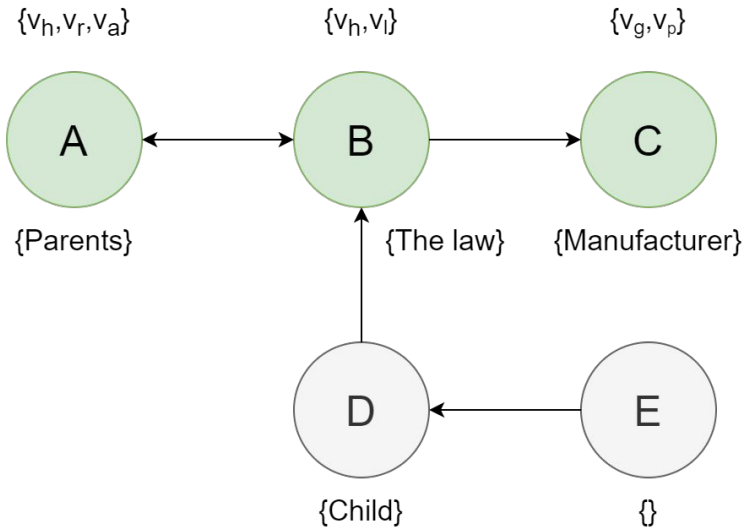


Abstract Argumentation Framework (AAF)

- $F_v = (A_p, A_e, R, A_g, V, val, \pi)$
- A_p: Set of practical arguments
- A_e: Set of epistemic arguments
- $R \subseteq (A_p x A_p) \cup (A_e x A_e) \cup (A_p x A_e)$
- Ag: Set of agents (Stakeholders)
- V: Set of values
- $val: A_p \rightarrow 2^V$
- $\pi: A_p \cup A_e \rightarrow 2^{Ag}$
- $F_v = (A_p \cup A_e, R)$ (reduced form)

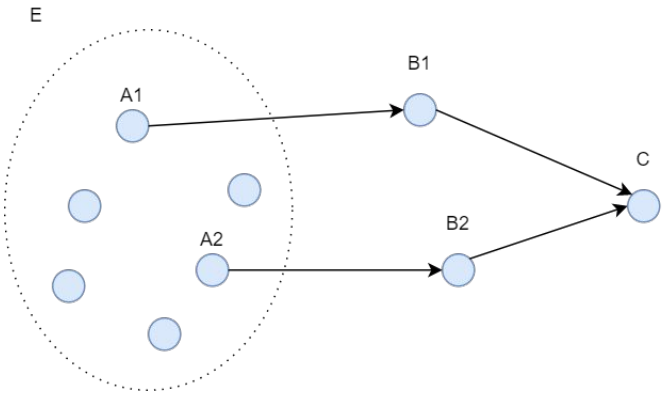
Abstract Argumentation Framework





Terminology

- A set of arguments that can be accepted together called ***extension***
- E is ***conflict-free*** iff E does not contain A, B, such that A attacks B
- E ***defends*** an argument C iff for each argument B that attacks C, E contains an argument that attacks B



Terminology contd.

E is:

- ***admissible*** iff it is conflict-free and legal labelling w.r.t in/out
- ***complete extension*** iff E is admissible and legal labelling w.r.t in/out/undec
- ***preferred extension*** iff E is a maximal complete extension (w.r.t set inclusion)
- ***grounded extension*** iff E is a minimal complete extension (w.r.t set inclusion)

Agreement reaching

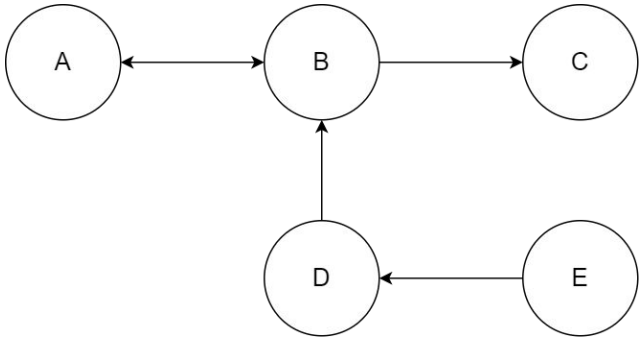
Step 1:

- Compute set of extensions in a reduced AAF

Step 2:

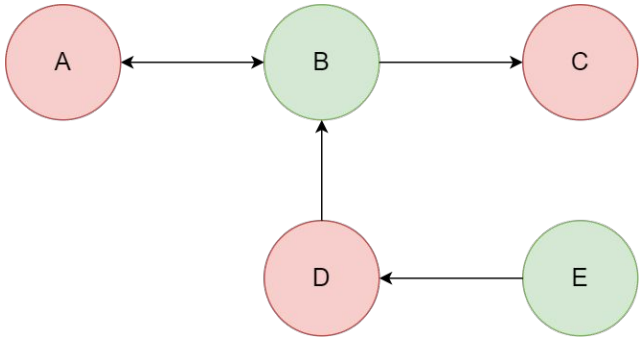
- Choose a subset of extensions that maximizes the extent of agreement over V

Step 1: Compute set of extensions in a reduced AAF

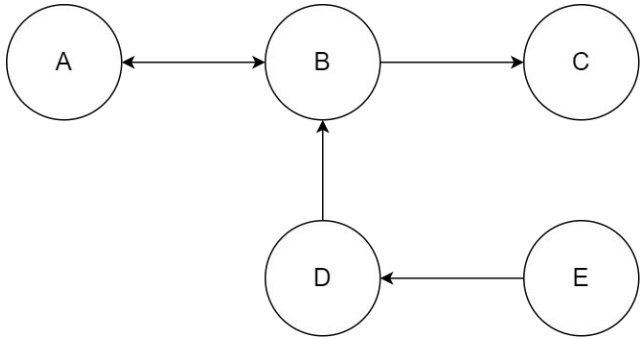


- Apply Argument Labelling
- in → all attackers are out
- out → there is an attacker that is in
- undec → not all attackers are out and no attacker is in

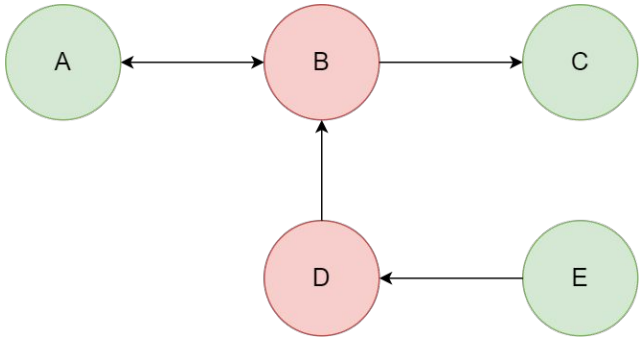
E1 = {B, E}



Step 1 contd.

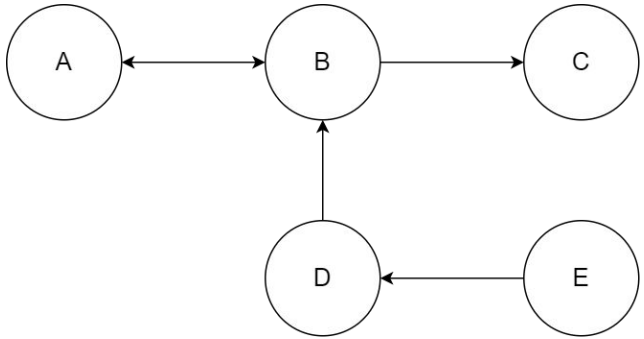


E2 = {A, C, E}



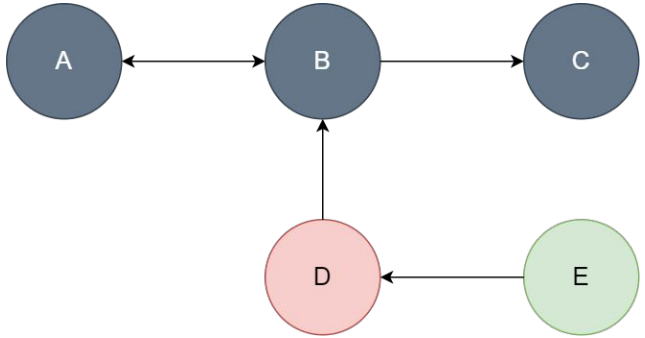
- Apply Argument Labelling
- in → all attackers are out
- out → there is an attacker that is in
- undec → not all attackers are out and no attacker is in

Step 1 contd.

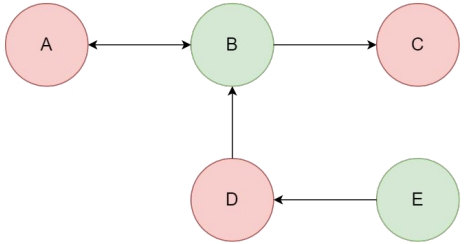


- Apply Argument Labelling
- in → all attackers are out
- out → there is an attacker that is in
- undec → not all attackers are out and no attacker is in

E3 = {E}

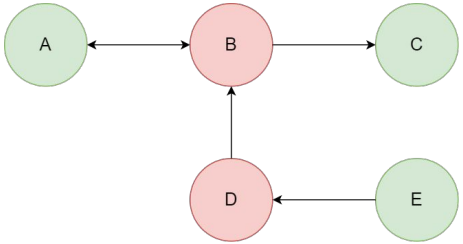


Step 1 contd.



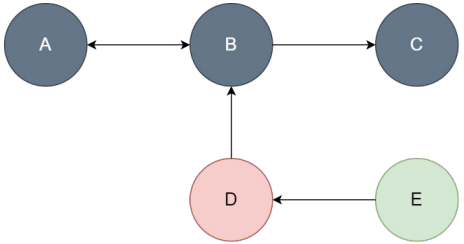
in = {B,E} out={A,C,D} undec={}

Preferred extension



in = {A,C,E} out={B,D} undec={}

Preferred extension



in = {E} out={D} undec={A,B,C}

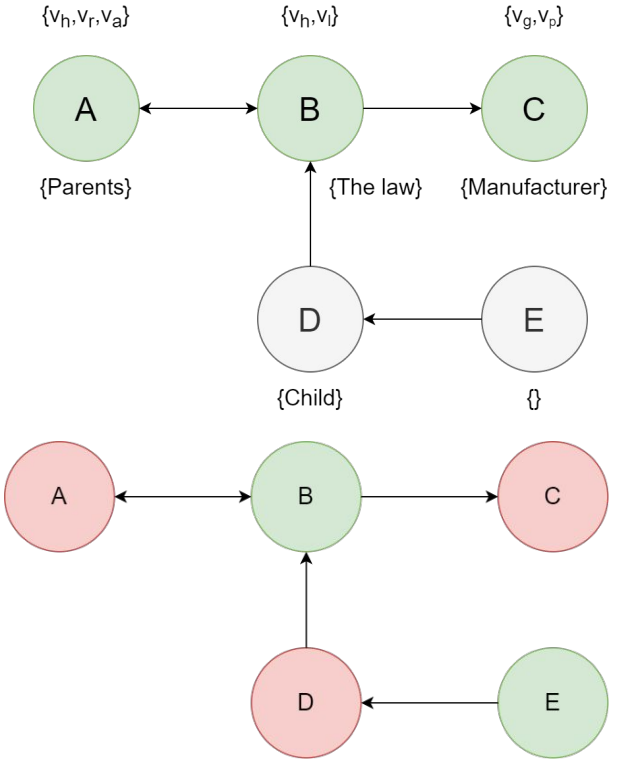
Grounded extension

Step 2: Maximize the extent of agreement over V

- Let E, E’ $\subseteq$ A
- $V_E = U_{A \in E \cap Ap} val(A)$
- $V_{E'} = U_{A \in E' \cap Ap} val(A)$
- V_E reaches maximal extent of agreement over V iff $\nexists$E’...
- … such that $V_{E'} > V_E$ (in terms of priority)

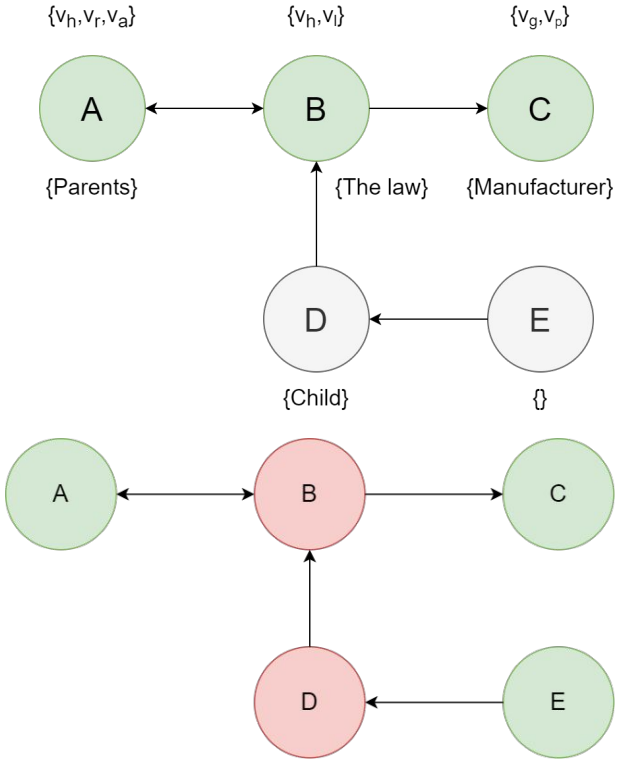
Example - What is V_E

- $V_E = U_{A \in E \cap Ap} val(A)$
- E1 = {B, E}
- A_p = {A, B, C}
- $V_{E1} = val(B) = \{v_h, v_l\}$



Example - What is V_E (contd.)

- $V_E = U_{A \in E \cap Ap} val(A)$
- E2 = {A, C, E}
- A_p = {A, B, C}
- V_{E2} = val(A) U val(C) = $\{v_h, v_r, v_a, v_g, v_p\}$



Step 2: Lifting Principle

- Elitist principle $V_1 \succeq_{Eli} V_2$ iff:
  - $\forall\ v' \in V_1\ \exists\ v \in V_2$ such that $v' \geq v$

- Democratic principle $V_1 \succeq_{Dem} V_2$ iff:
  - $\exists\ v' \in V_1\ \forall\ v \in V_2$ such that $v' \geq v$

Step 2: Lifting Principle - Example

Consider following partial ordering:

- $v_l \geq v_r \geq v_p \geq v_a \geq v_g \geq v_h$
- $V_{E1} = \{v_h, v_l\}$
- $V_{E2} = \{v_h, v_r, v_a, v_g, v_p\}$

- Elitist principle $V_1 \succeq_{Eli} V_2$ iff:
  - $\forall\ v' \in V_1\ \exists\ v \in V_2$ such that $v' \geq v$
- Democratic principle $V_1 \succeq_{Dem} V_2$ iff:
  - $\forall\ v \in V_2 \exists\ v' \in V_1$ such that $v' \geq v$

- We have $V_{E1} \succeq_{Eli} V_{E2}$ since $v_h \geq v_h$ and $v_l \geq v_h$
- We have $V_{E1} \succeq_{Dem} V_{E2}$ since $v_l \geq v_{h,r,a,g,p}$

Example contd.

Consider following partial ordering:

- $v_l \geq v_r \geq v_p \geq v_a \geq v_g \geq v_h$
- $V_{E1} = \{v_h, v_l\}$
- $V_{E2} = \{v_h, v_r, v_a, v_g, v_p\}$

- Elitist principle $V_1 \succeq_{Eli} V_2$ iff:
  - $\forall\ v' \in V_1\ \exists\ v \in V_2$ such that $v' \geq v$
- Democratic principle $V_1 \succeq_{Dem} V_2$ iff:
  - $\forall\ v \in V_2 \exists\ v' \in V_1$ such that $v' \geq v$

- We have $V_{E2} \succeq_{Eli} V_{E1}$ since $v_{h,r,a,g,p} \geq v_h$
- We do not have $V_{E2} \succeq_{Dem} V_{E1}$ since $\nexists v \in V_2\ v \geq v_l$

Example Lifting - Conclusion

- E1 maximizes the extent of agreement over the set of values by using both the democratic and elitist principles. Since we have:
  - $V_{E1} \succeq_{Eli} V_{E2}$
  - $V_{E1} \succeq_{Dem} V_{E2}$

- E2 maximizes the extent of agreement over the set of values by using the elitist principle. Since we have:
  - $V_{E2} \succeq_{Eli} V_{E1}$

Agreement Reaching

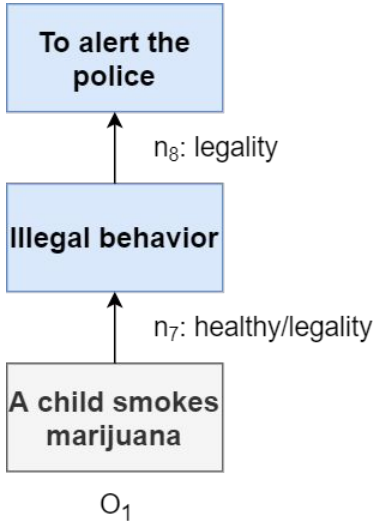
- Derivability
- Agreement Reaching
- Justification in a Dialogue Graph

Agreement Reaching

- Assume → maximize the extent of agreement using the democratic principle
- Action to be selected → Alert the police
- Because...

Derivability

“**The police should be alerted**” is a conclusion of an argument B, which can be derived from an observation “a child smokes marijuana” and two norms “if a child smokes marijuana, their behavior counts as an illegal behavior” and “if there is an illegal behavior then the police should be alerted”.

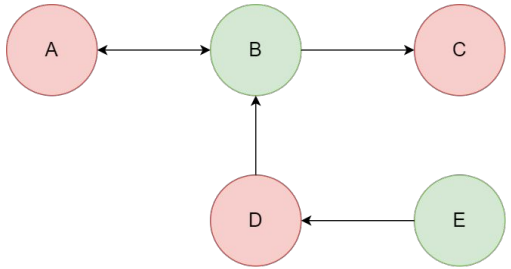


Agreement Reaching

The extension E1 = {B, E} which contains the argument B is selected since E1 maximizes the extent of agreement over the set of values by using democratic principle.

Justification in a Dialogue Graph - Discussion Game

- Play rules:
  - Every move of M (besides the first one) needs to be an attacker of the directly preceding move of S
  - Every move of S needs to be attacker of some previous move of M
  - S is not allowed to repeat his moves
  - M can repeat his moves
- Winning rules:
  - If S uses an argument that was previously used by M then S wins
  - If M uses an argument that was previously used by S then S wins
  - If M cannot make a move then S wins
  - If S cannot make a move then M wins



M: in(B)
S: out(A)
M: in(B)
S: out(D)
M: in(E)

M wins the game, S can not move

Assumptions

- Clarify origin and priority of the values for an AMA
- Knowledge-based representation
- Stakeholder assumptions
- Argumentation-based engine assumptions

Challenges

- How to decide on the ordering of the ethical values?
- How to ensure that all stakeholders are treated fairly?

Conclusion

- Argumentation-based architecture
- Moral agents as social agents
- Ability to take reasoning of others into account…
- … by combining normative systems of multiple stakeholders to…
- … reach an ethical decision.

Thank you for listening.