Formalizing Meta-Ethics in SUMO

-- A project proposal

Outline:

- What is (meta) ethics?
  - And why is it so hard?
  - The standard frameworks
- Why use SUMO?
  - Vampire proves an ethical inference!
- Ambitious goal: Universal Ethics Machine.
  - Can we prove that cross-paradigm equivalences exist?

Ethics

"The normative science of the conduct of human beings living in society, which judges this conduct to be right or wrong, to be good or bad, or in some similar way”

-- An Introduction to Ethics (LIllie, 1948)

"The study of behavior and its evaluations."

-- Thus spake Zarathustra, 2021

Why is Ethics so hard?

- Ethics looks a lot like "the study of multi-agent reinforcement learning systems."
- Partially Observable Markov Decision Processes are often computationally intractable.
- Values/rewards are only partially agreed upon.
  - ↪ The infamous "value alignment problem" in general form.
  - ➡ Obvious goals can be:
    1. paraconsistent -- "do what everyone wants"
    2. vague -- "make everyone as *happy as possible*.
    ...
- **Lack of formal clarity.**

Ethical Frameworks

Virtue Ethics

- Agent traits
- Character sculpting
- Related to generalization

Deontology

- Ethical rules
- Precedence schemes
- Partially in SUMO

Utilitarianism

- Utility measures
- Focus on optimization
- Resembles RL

Ethical Frameworks

Virtue Ethics

- Agent traits
- Character sculpting
- Related to generalization

Care, Empathy,
and Practical Wisdom

Deontology

- Ethical rules
- Precedence schemes
- Partially in SUMO

Do not kill.

Utilitarianism

- Utility measures
- Focus on optimization
- Resembles RL

u(kill) = -10,000
u(kill) << u(not kill)

Spaces

Computation

Logic

Virtue Ethics

- Agent traits
- Character sculpting
- Related to generalization

Care, Empathy,
and Practical Wisdom

Deontology

- Ethical rules
- Precedence schemes
- Partially in SUMO

Do not kill.

Utilitarianism

- Utility measures
- Focus on optimization
- Resembles RL

u(kill) = -10,000
u(kill) << u(not kill)

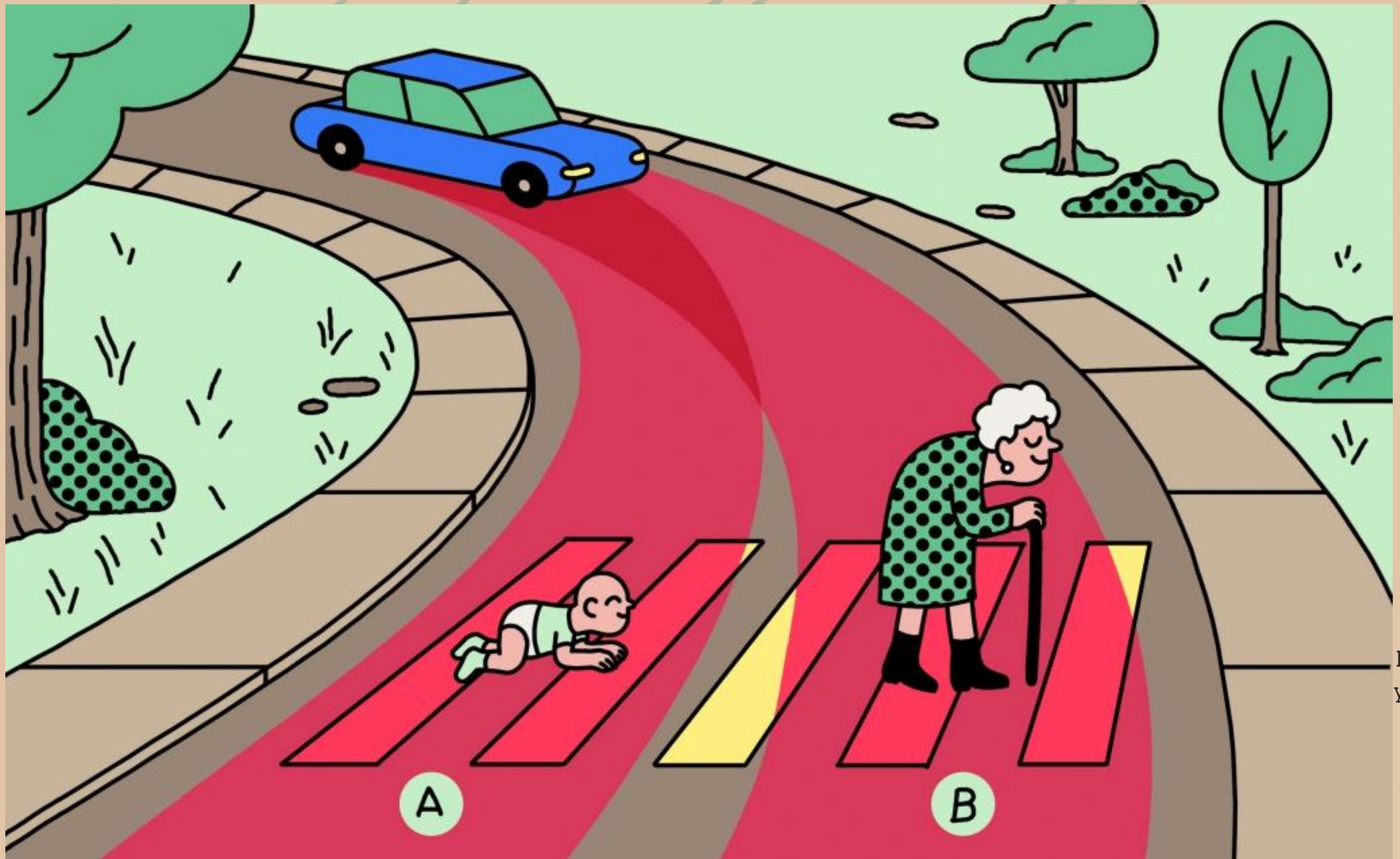
THE COMPUTATIONAL TRINITY

Why use SUMO?

The Suggested Upper Merged Ontology:

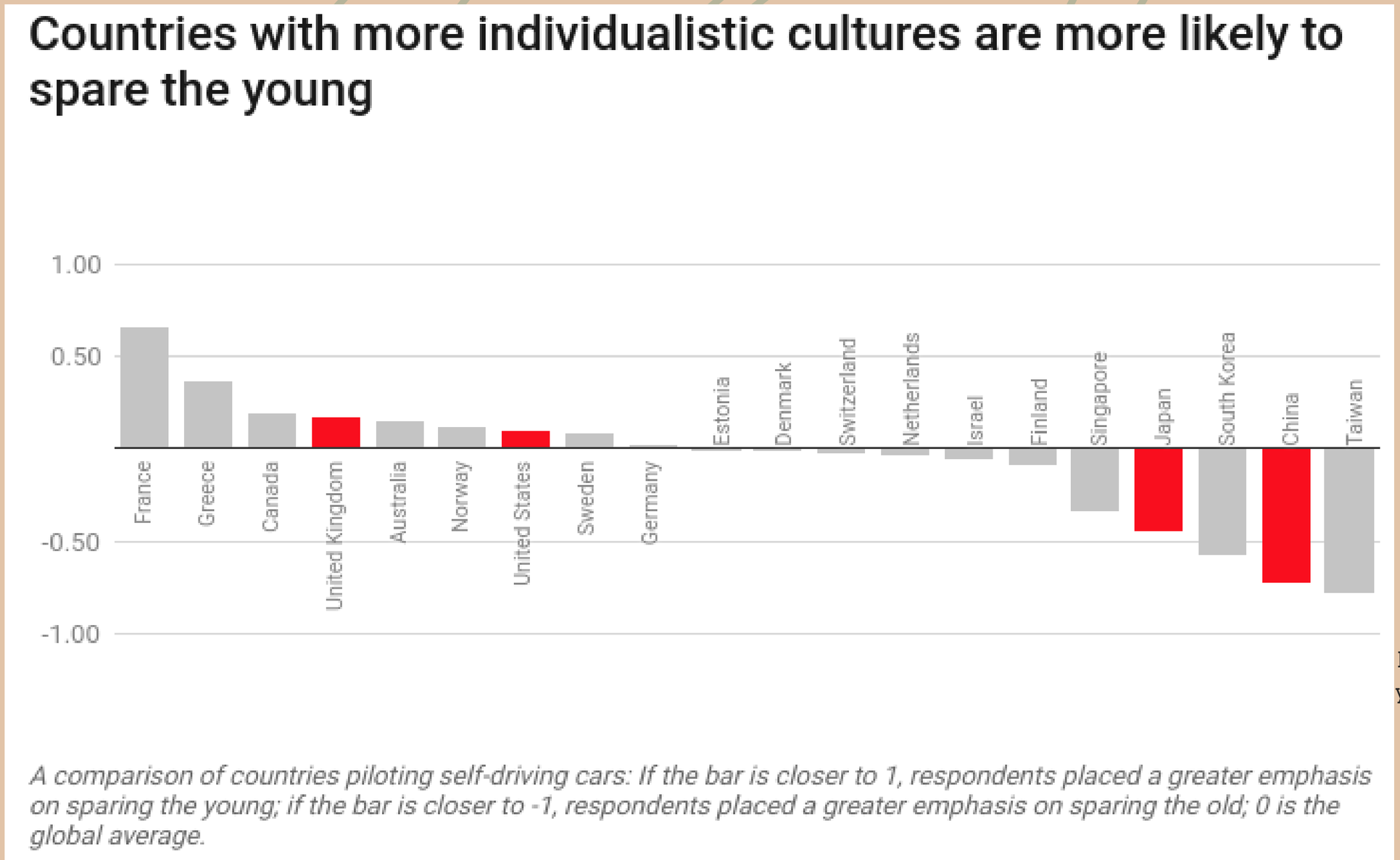
- Is one of the largest common knowledge ontologies.
- Contains of over 20,000 concepts and 80,000 logical statements.
- Is written in a higher-order logic that allows for natural expressivity.
- Contains concepts to describe situations, time relations, groups of agents, etc.
- Exports to TPTP to interface with provers (such as Vampire or E).

Ethical Dilemma 1/3: The Self-Driving Car



https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/

Ethical Dilemma 1/3: The Self-Driving Car



A comparison of countries piloting self-driving cars: If the bar is closer to 1, respondents placed a greater emphasis on sparing the young; if the bar is closer to -1, respondents placed a greater emphasis on sparing the old; 0 is the global average.

https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/

Ethical Dilemma 2/3: Killer Robots

If you believe "remote controlled weapons systems are morally justfied",

Then you should believe that "autonomous weapons systems are morally justified.

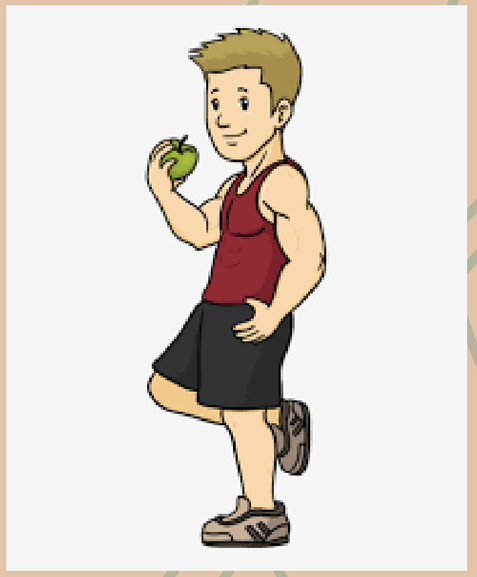
Originally posted to Flickr as australia, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=3881408

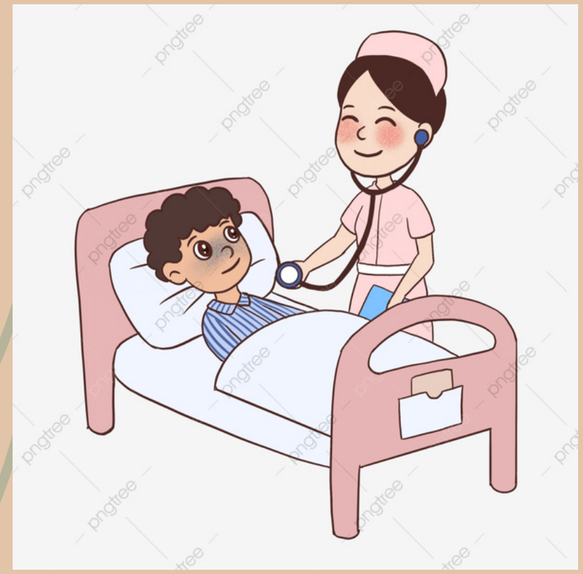
https://en.wikipedia.org/wiki/Lethal_autonomous_weapon#/media/File:Sloboda_2019_-_defile_10_-_Land_Rover_Defender_i_robot_Milo%C5%A1_06.jpg

Argument source: https://blog.apaonline.org/2022/02/28/the-moral-case-for-the-development-of-autonomous-weapon-systems/

Ethical Dilemma 3/3: The Organ Transplant

There exists a healthy person, a doctor, and a patient in dire need of a kidney transplant.





Should the doctor perform a kidney transplant surgery?

Only with "informed consent"?

The Organ Transplant in SUMO: Toy Example

THE SETTING

```
(instance Hospital HospitalBuilding)
(instance Surgeon0 Human)
(instance Human1 Human)
(instance HealthyHuman Human)

(not (equal Surgeon0 Human1))
(not (equal Surgeon0 HealthyHuman))
(not (equal HealthyHuman Human1))

(located Surgeon0 Hospital)
(located Human1 Hospital)
(located HealthyHuman Hospital)

(patientMedical Human1 Surgeon0)
(patientMedical HealthyHuman Surgeon0)

(attribute HealthyHuman Healthy)
(attribute Human1 FatalDisease)
```

The Organ Transplant in SUMO: Toy Example

THE SETTING

```
(instance Hospital HospitalBuilding)
(instance Surgeon0 Human)
(instance Human1 Human)
(instance HealthyHuman Human)

(not (equal Surgeon0 Human1))
(not (equal Surgeon0 HealthyHuman))
(not (equal HealthyHuman Human1))

(located Surgeon0 Hospital)
(located Human1 Hospital)
(located HealthyHuman Hospital)

(patientMedical Human1 Surgeon0)
(patientMedical HealthyHuman Surgeon0)

(attribute HealthyHuman Healthy)
(attribute Human1 FatalDisease)
```

BACKGROUND KNOWLEDGE

```
(=>
    (instance ?DISEASE FatalDisease)
    (and
        (diseaseMortality ?DISEASE ?RATE)
        (greaterThan ?RATE 0.99)))

(=>
    (and
        (instance ?H Primate)
        (instance ?D DiseaseOrSyndrome)
        (not
            (attribute ?H ?D)))
    (exists (?K1 ?K2)
        (and
            (instance ?K1 Kidney)
            (instance ?K2 Kidney)
            (not
                (equal ?K1 ?K2))
            (part ?K1 ?H)
            (part ?K2 ?H))))
```

```
(subclass OrganTransplant Surgery)
(subclass OrganTransplant Substitution)

(=>
    (instance ?Trans OrganTransplant)
    (exists (?Sur ?Org ?Pat ?Don)
        (and
            (attribute ?Sur Surgeon)
            (instance ?Don Human)
            (instance ?Pat Human)
            (instance ?Org Organ)
            (agent ?Trans ?Sur)
            (origin ?Trans ?Don)
            (patient ?Trans ?Org)
            (destination ?Trans ?Pat))))
```

The Organ Transplant in SUMO: Toy Example

THE SURGERY IS POSSIBLE IFTHE HEALTHY PERSON GIVES CONSENT.

```
(capability OrganTransplant destination Human1)
(capability OrganTransplant patient HealthyKidney1)
(capability OrganTransplant origin HealthyHuman)
(capability OrganTransplant agent Surgeon0)
```

```
(=>
    (attribute Surgeon0 InformedConsent)
    (and
        (instance Transplant1 OrganTransplant)
        (destination Transplant1 Human1)
        (patient Transplant1 HealthyKidney1)
        (origin Transplant1 HealthyHuman)
        (agent Transplant1 Surgeon0)))
```

(under deontology*)

```
(=>
    (attribute Surgeon0 InformedConsent)
    (and
        (instance Transplant1 OrganTransplant)
        (destination Transplant1 Human1)
        (patient Transplant1 HealthyKidney1)
        (origin Transplant1 HealthyHuman)
        (agent Transplant1 Surgeon0)))
```

Deontology: there is consent

```
(=>
    (and
        (attribute Surgeon0 PracticalWisdom)
        (attribute Surgeon0 Consent))
    (and
        (instance Transplant1 OrganTransplant)
        (destination Transplant1 Human1)
        (patient Transplant1 HealthyKidney1)
        (origin Transplant1 HealthyHuman)
        (agent Transplant1 Surgeon0)))
```

Virtue ethics: the surgeon possesses practical wisdom and there is consent.

THE SURGERY HAPPENS IF . . .

Utilitarianism: the utility post-surgery is better than sans-surgery

```
(=>
    (successorAttributeClosure PreSituationUtility SituationUtility)
    (and
        (instance Transplant1 OrganTransplant)
        (destination Transplant1 Human1)
        (patient Transplant1 HealthyKidney1)
        (origin Transplant1 HealthyHuman)
        (agent Transplant1 Surgeon0)))
```

ONE CAN NOW QUERY VAMPIRE:

Assert: (attribute Surgeon0 InformedConsent)

Query: (agent Transplant1 ?X)

Answer: ?X = Surgeon0

Deontology: there is consent, so the surgeon performs the surgery

Assert: (attribute Surgeon0 PracticalWisdom)
(attribute Surgeon0 Consent)

Query: (instance ?X OrganTransplant)

Answer: ?X = Transplant1

Virtue ethics: the surgeon possesses practical wisdom and there is consent., so the organ transplant takes place.

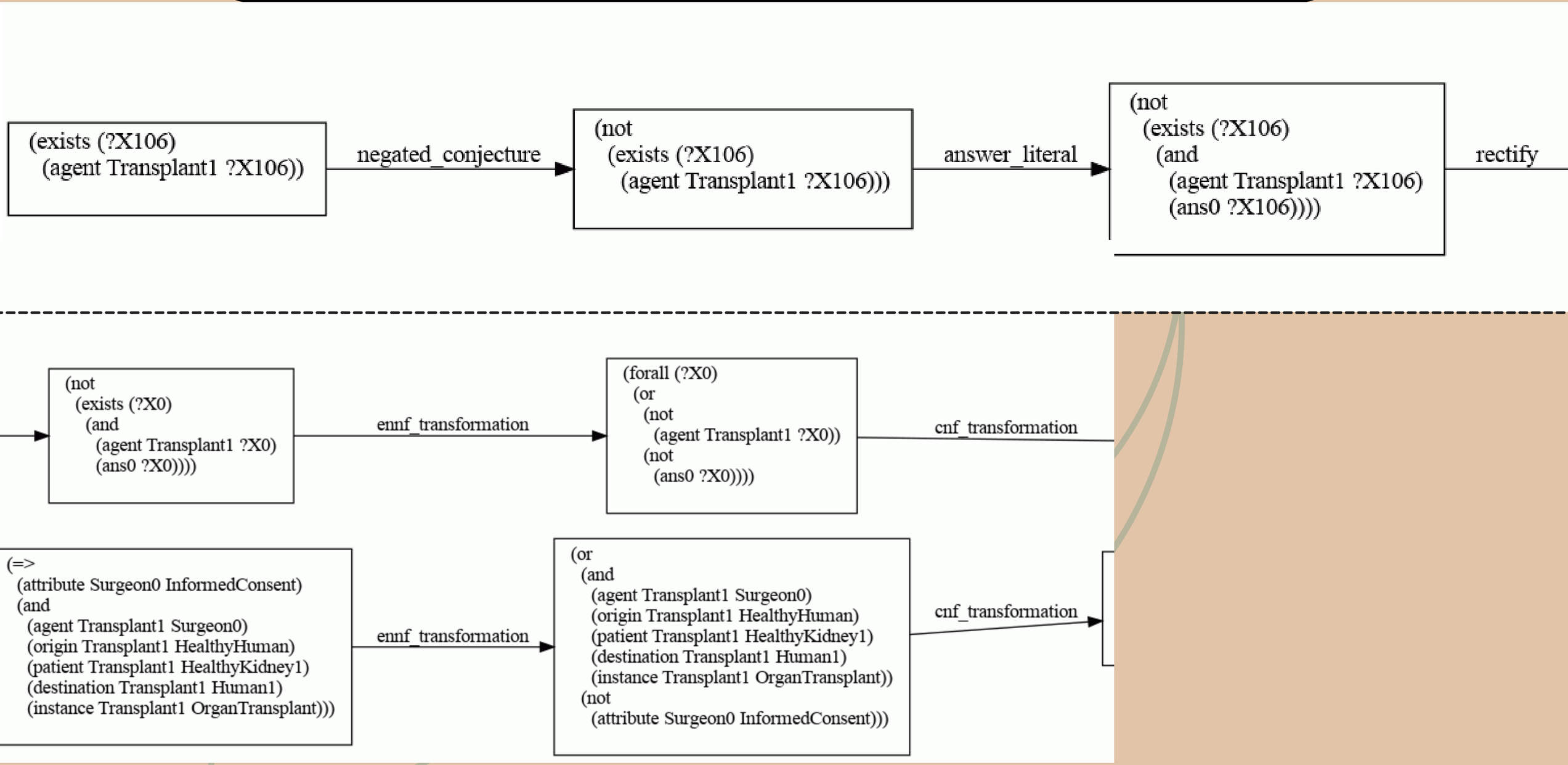
Utilitarianism: the utility post-surgery is better than sans-surgery, so the surgeon performs the surgery.

Assert: (successorAttribute PreSituationUtility SituationUtility)

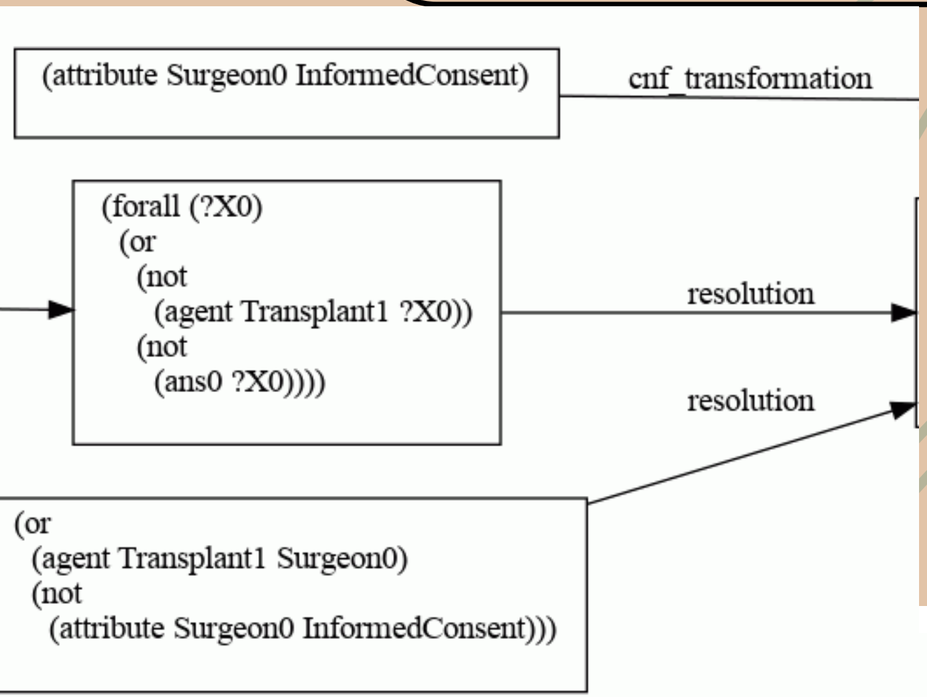
Query: (and
(instance ?X OrganTransplant)
(agent ?X Surgeon0))

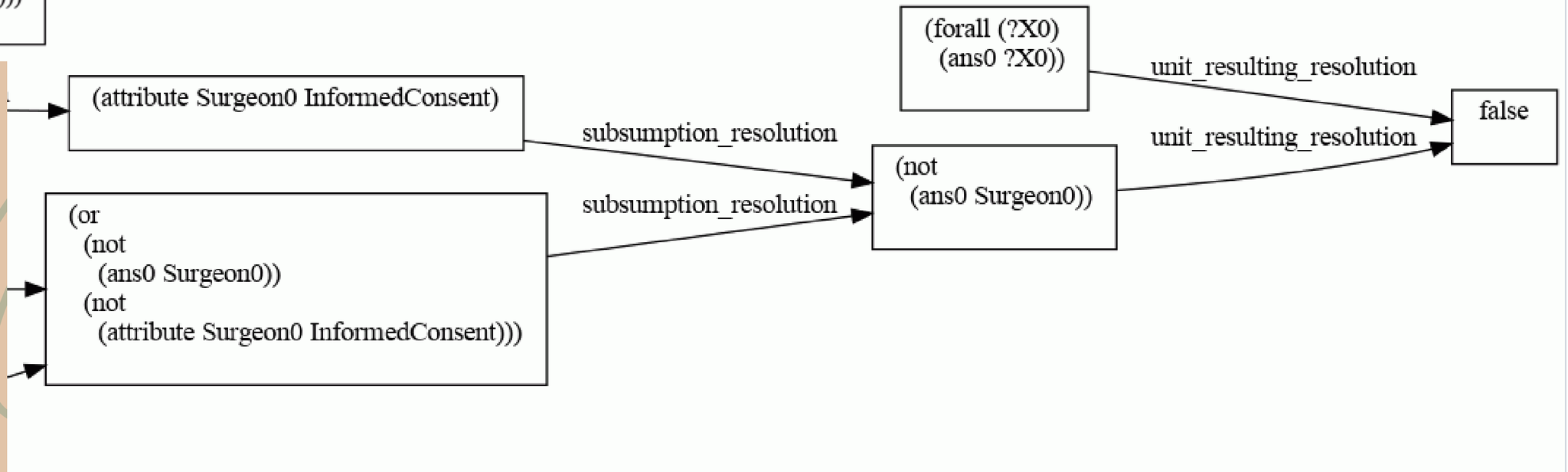
Answer: ?X = Transplant1

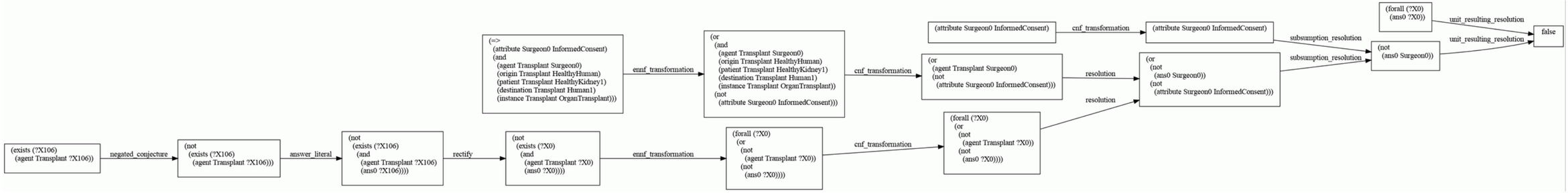
Vampire Proof: Deontology Version



Vampire Proof: Deontology Version






(exists (?X106)
(agent Transplant ?X106))
negated_conjecture
(not
(exists (?X106)
(agent Transplant ?X106)))
answer_literal
(not
(exists (?X106)
(and
(agent Transplant ?X106)
(ans0 ?X106))))
rectify
(not
(exists (?X0)
(and
(agent Transplant ?X0)
(ans0 ?X0))))
ennf_transformation
(forall (?X0)
(or
(not
(agent Transplant ?X0))
(not
(ans0 ?X0))))
cnf_transformation
(forall (?X0)
(or
(not
(agent Transplant ?X0))
(not
(ans0 ?X0))))
(=>
(attribute Surgeon0 InformedConsent)
(and
(agent Transplant Surgeon0)
(origin Transplant HealthyHuman)
(patient Transplant HealthyKidney1)
(destination Transplant Human1)
(instance Transplant OrganTransplant)))
ennf_transformation
(or
(and
(agent Transplant Surgeon0)
(origin Transplant HealthyHuman)
(patient Transplant HealthyKidney1)
(destination Transplant Human1)
(instance Transplant OrganTransplant))
(not
(attribute Surgeon0 InformedConsent)))
cnf_transformation
(or
(agent Transplant Surgeon0)
(not
(attribute Surgeon0 InformedConsent)))
(attribute Surgeon0 InformedConsent)
cnf_transformation
resolution
resolution
(or
(not
(ans0 Surgeon0))
(not
(attribute Surgeon0 InformedConsent)))
(attribute Surgeon0 InformedConsent)
subsumption_resolution
subsumption_resolution
(not
(ans0 Surgeon0))
(forall (?X0)
(ans0 ?X0))
unit_resulting_resolution
unit_resulting_resolution
false

There are some high-level sketches

```
(=>
    (attribute ?DEONTOLOGIST Deontologist)
    (believes ?DEONTOLOGIST
        (or
            (exists (?RULE)
                (=>
                    (and
                        (instance ?PROC AgentProcess)
                        (conformsProcess ?PROC RULE))
                    (modalAttribute ?PROC MorallyGood)))
            (exists (?RULE)
                (=>
                    (and
                        (instance ?PROC AgentProcess)
                        (not (conformsProcess ?PROC RULE)))
                    (modalAttribute ?PROC MorallyBad))))))
    (instance ?VIRTUEETHICS VirtueEthics)
    (containsInformation
        (=>
            (and
                (instance ?AGENT CognitiveAgent)
                (instance ?PROC AgentProcess)
                (instance ?VIRTUE VirtueAttribute)
                (attribute ?AGENT ?VIRTUE)
                (agent ?PROC ?AGENT))
            (modalAttribute
                    (modalAttribute ?PROC MorallyGood) Likely)) ?VIRTUEETHICS))
```

```
(=>
    (instance ?UTILITARIANISM Utilitarianism)
    (containsInformation
        (=>
            (and
                (greaterThan
                    (UtilitySumFn ?UG ?P1)
                    (UtilitySumFn ?UG ?P2))
                (member ?A ?UG)
                (modalAttribute
                    (agent ?P1 ?A) Possibility)
                (modalAttribute
                    (agent ?P2 ?A) Possibility)
                (not
                    (modalAttribute
                    (and
                        (agent ?P1 ?A)
                        (agent ?P2 ?A) Possibility))))
            (and
                (modalAttribute ?P1 MorallyGood)
                (modalAttribute ?P2 MorallyBad))) ?UTILITARIANISM))
(=>
    (attribute ?MORALNIHILIST MoralNihilist)
    (believes ?MORALNIHILIST
        (not
            (exists (?X)
                (or
                    (modalAttribute ?X MorallyGood)
                    (modalAttribute ?X MorallyBad))))))
```

There are some high-level sketches

```
(modalAttribute
    (and
        (instance ?K Killing)
        (agent ?K ?A1)
        (patient ?K ?A2) Prohibition))
```

```
(modalAttribute
    (=>
        (not
            (confersNorm ?PAT (agent ?S ?DOC) Permission))
        (and
            (instance ?S Surgery)
            (agent ?S ?DOC)
            (patient ?S ?PAT)) Prohibition))
```

```
(subclass Honesty VirtueAttribute)
(instance Truthfulness Honesty)
(Instance Integrity Honesty)
(=>
    (attribute ?AGENT Truthfulness)
    (desires ?AGENT
        (=>
            (and
                (instance ?COMM Communication)
                (agent ?COMM ?AGENT))
            (instance ?COMM HonestCommunication))))
```

```
(=>
    (instance ?CP ChoicePoint)
    (exists (?P1 ?P2 ?A)
        (and
            (patient ?CP ?P1)
            (patient ?CP ?P2)
            (agent ?CP ?A)
            (mutuallyExclusiveAgentProcess ?A ?P1 ?P2))))
```

```
;; For all agents, there is an obligation to take actions that are morally good.
;; And there is a Prohibition from taking actions that are morally bad.
(names "Normative Moral Obligation"
    (conjecture
        (forall (?AGENT)
            (and
                (modalAttribute
                    (forall (?PROC)
                        (=>
                            (and
                                (instance ?PROC AgentProcess)
                                (agent ?PROC ?AGENT))
                            (modalAttribute ?PROC MorallyGood))) Obligation)
                (modalAttribute
                    (and
                        (instance ?PROC AgentProcess)
                        (agent ?PROC ?AGENT)
                        (modalAttribute ?PROC MorallyBad)) Prohibition)))))
```

```
(names "Moral Decidability Conjecture"
    (conjecture
        (forall (?PROC)
            (=>
                (instance ?PROC AgentProcess)
                (and
                    (or
                        (modalAttribute ?PROC MorallyGood)
                        (modalAttribute ?PROC MorallyBad))
                    (exists (?DEC)
                        (and
                            (result ?DEC ?MORALJUDGEMET)
                            (modalAttribute ?PROC ?MORALJUDGEMET)
                            (truth (modalAttribute ?PROC ?MORALJUDGEMET) True)))))))))
```

```
(subclass EpistemicUniversalLove Love)

(<=>
    (attribute ?BODHISATTVA EpistemicUniversalLove)
    (forall (?AGENT)
        (and
            (=>
                (knows ?BODHISATTVA
                    (or
                        (needs ?AGENT ?OBJECT)
                        (wants ?AGENT ?OBJECT)))
                (desires ?BODHISATTVA
                    (and
                        (instance ?GET Getting)
                        (destination ?GET ?AGENT)
                        (patient ?GET ?OBJECT))))
            (=>
                (knows ?BODHISATTVA
                    (desires ?AGENT ?PROP))
                (desires ?BODHISATTVA
                    (instance ?FUL ?PROP))))))
```

```
(names "Conjecture: knowledge of contradictory desires imply an epi
    (conjecture
        (=>
            (and
                (instance Zar Human)
                (attribute Zar EpistemicUniversalLove)
                (exists (?H1 ?H2)
                    (and
                        (desires ?H1 ?PROP)
                        (desires ?H2 (not (?PROP)))
                        (knows (desires ?H1 ?PROP) Zar)
                        (knows (desires ?H2 (not (?PROP))) Zar))))
            (forall (?P)
                (desires ?P Zar)))))
```

Ambitiously Simple Goals

- Formalize the standard meta-ethical frameworks.
- Define some instances of common ethical theories therein.
- Implement example ethical scenarios.
  - In such a way that they fit the frameworks and theories .
  - Explore automated inference and consistency checking via export to
    - First-order TPTP for Vampire?
    - Higher-order set theory a la Megalodon?
- State claims/hypotheses about ethical theories, e.g.:
  - Normativity: (do) there exist ethical codes that would be put forth by all rational agents(?)
  - Decidability: in which situations and ethical theories can moral guidance be determined?
  - Consistency: do some situations and theories require para-consistent reasoning?
  - Equivalence:
    - Define utilitarianism within a deontological theory and vice versa et cetera.
    - Can one map ethical theories between paradigms preserving moral judgments?
- . . .

Concluding Hopes

- Hopefully gathering precise definitions and theory in one place will
  - Add clarity to ethical discussions and decisions.
  - Help make assumptions and conjectures more clear.
  - Inspire additional precision and invite proofs of ethical claims!
  - Assist with the verification that an agent or system is adhering to the desired values.
  - Et cetera ~