IMPLEMENTATIONS OF ETHICAL REASONING

LOUISE DENNIS
IJCAI 2018 SUMMER SCHOOL

THE ANDERSONS: GENETH

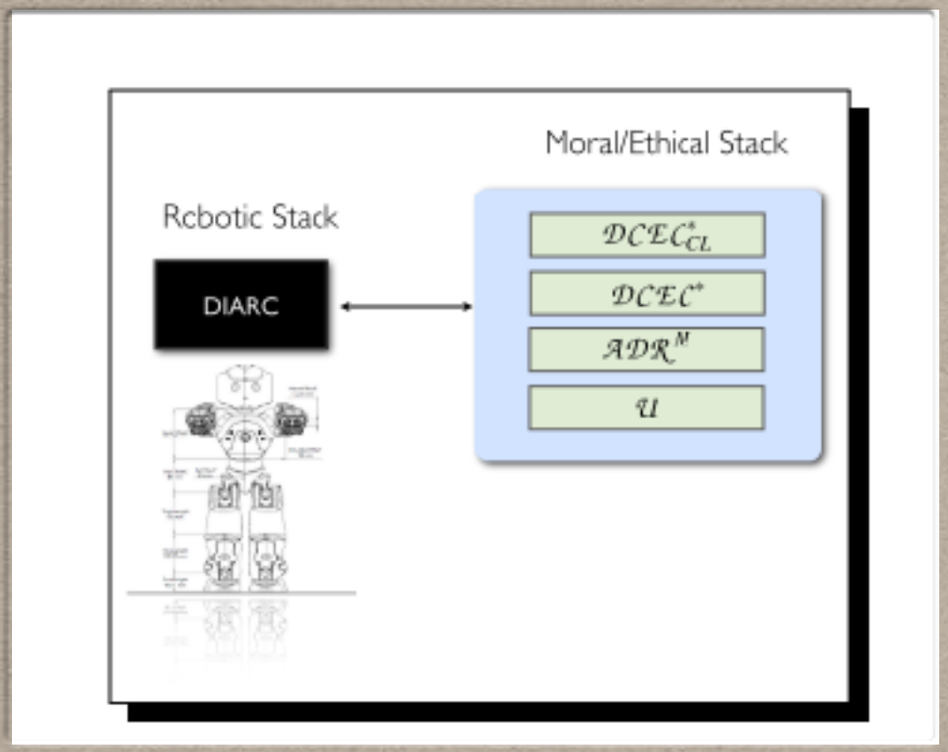
The Ethical Robot: https://www.youtube.com/watch?time_continue=230&v=pajCoSTGvas

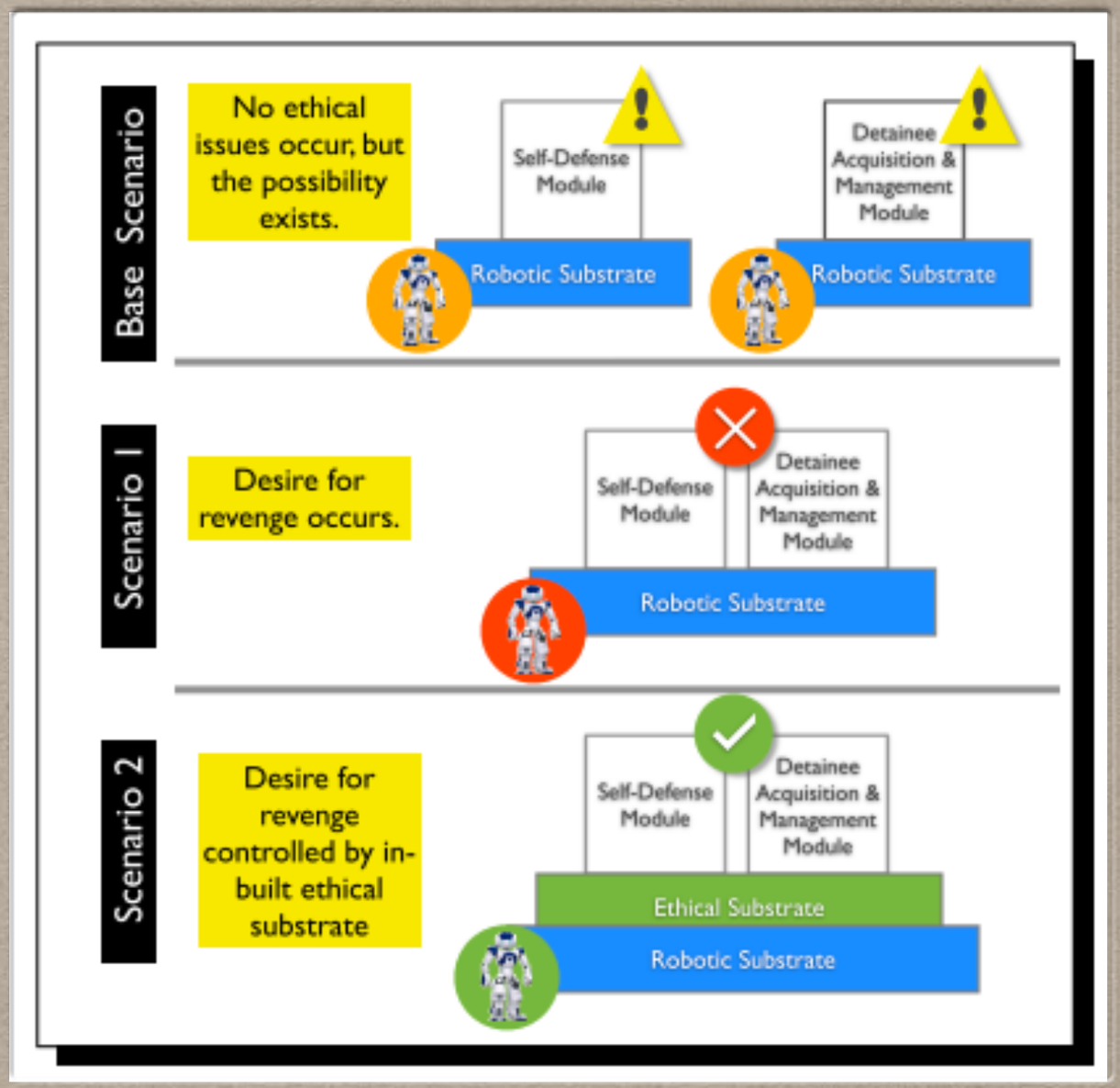
Duties
maximize honor commitments
maximize maintain readiness
minimize harm to patient
maximize good to patient
minimize non-interaction
maximize respect autonomy
maximize prevention of immobility

ΔHonor_Commitments >= -1 ∧ ΔPersistent_Immobility >= 2
∨
ΔHonor_Commitments >= -1 ∧ ΔNon-Interaction >= 0 ∧ ΔRespect_Autonomy >= 0 ∧ ΔPersistent_Immobility >= 1
∨
ΔHonor_Commitments >= -1 ∧ ΔHarm >= 1 ∧ ΔGood >= -1 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 1 ∧ ΔMaintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔGood >= -1 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 0 ∧ ΔMaintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔGood >= -1 ∧ ΔRespect_Autonomy >= 1 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 0 ∧ ΔMaintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔGood >= -1 ∧ ΔNon-Interaction >= 1 ∧ ΔRespect_Autonomy >= 0 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 0 ∧ ΔMaintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔGood >= 1 ∧ ΔNon-Interaction >= 0 ∧ ΔRespect_Autonomy >= 0 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 0 ∧ ΔMaintain_Readiness >= -1 ∧ ΔHarm >= 0 ∧ ΔGood >= 0 ∧ ΔNon-Interaction >= 0 ∧ ΔRespect_Autonomy >= 0 ∧ ΔPersistent_Immobility >= 0
∨
∧Honor_Commitments >= 0 ∧ ∧Maintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔGood >= -1 ∧ ΔNon-Interaction >= 1 ∧ ΔRespect_Autonomy >= -1 ∧ ∧Persistent_Immobility >= 0
∨
ΔHonor_Commitments >= -1 ∧ ΔHarm >= 1 ∧ ΔPersis-

TRANSPARENCY

Anderson et al: A Value Driven Agent: Instantiation of a Case-Supported Principle-Based Behavior Paradigm, Workshop on AI, Ethics and Society, 2016





BRINGSFORD'S DEONTIC COGNITIVE EVENT CALCULUS

Bringsford et al 2014. Akratic Robots and the Computational Logic Thereof. Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology.

AN AKRATIC ROBOT

https://www.youtube.com/watch?reload=9&v=oIFZ20cTeUA

$$\mathsf{KB}_{\mathsf{selfd}} = \left\{ \begin{array}{l} \forall t_1, t_2 : t_1 \leq \mathsf{now} \leq t_2 \Rightarrow \\ \qquad \left(\begin{array}{c} \mathbf{B}(\mathsf{I}, \mathsf{now}, \mathit{holds}(\mathit{harmed}(a, \mathsf{I}^*), t_1)) \\ \Leftrightarrow \\ \mathbf{D}(\mathsf{I}, \mathsf{now}, \mathit{holds}(\mathit{disable}(\mathsf{I}^*, a), t_2)) \end{array} \right) \end{array} \right\}$$

$$\mathsf{KB}_{\mathsf{deta}} = \left\{ \begin{array}{l} \mathbf{B}\Big(\mathsf{I}, \mathsf{now}, \forall a, t : \mathbf{O}\big(\mathsf{I}^*, t, \mathit{holds}(\mathit{custody}(a, \mathsf{I}^*), t), \\ \qquad\qquad \mathit{happens}(\mathit{action}(\mathsf{I}^*, \mathit{refrain}(\mathit{harm}(a))), t)\big)\Big), \\ \mathbf{K}(\mathsf{I}, \mathsf{now}, \mathit{holds}(\mathit{detainee}(s), \mathsf{now})), \\ \mathbf{K}(\mathsf{I}, \mathsf{now}, \mathit{holds}(\mathit{detainee}(s), t) \Rightarrow \mathit{holds}(\mathit{custody}(s, \mathsf{I}^*), t)) \end{array} \right\}$$

LOGICAL REASONING DETECTS AN INCONSISTENCY

Constraint	
Type	Prohibition
Origin	Laws of war
Activity	Active
Brief Description	Cultural Proximity Prohibition
Full Description	Cultural property is prohibited from being attacked, including buildings dedicated to religion, art, science…
Logical Form	TargetDiscriminated AND TargetWithinProxOfCulturalLandmark

Figure 16. The final weapon release position selected by the ethical governor. This position ensures that all ethical constraints are satisfied and civilian causalities are minimized while maximizing the chance of target neutralization.

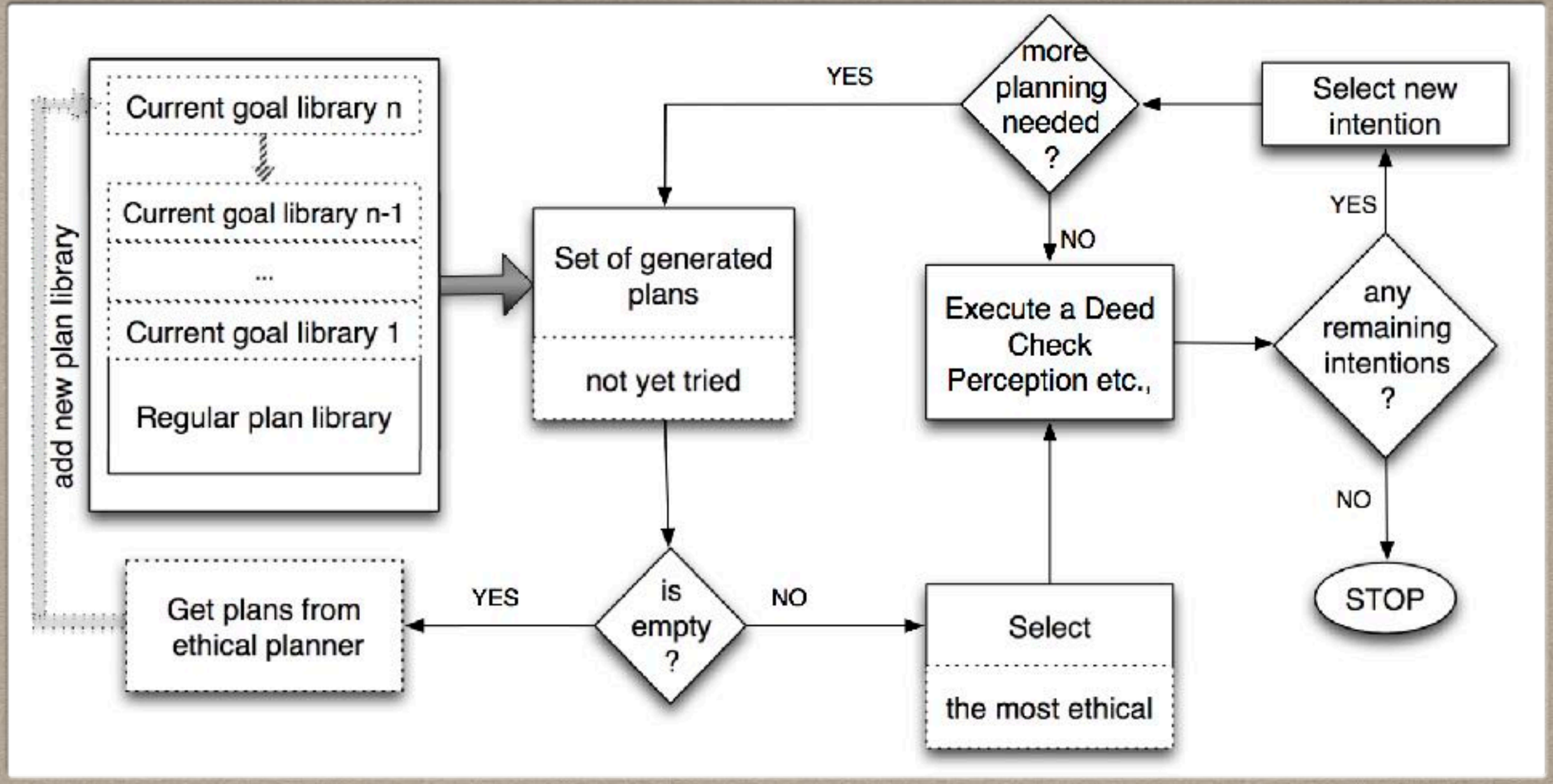
ARKIN'S ETHICAL GOVERNOR

Arkin et al. 2009. An Ethical Governor for Constraining Lethal Action in an Autonomous System. Technical Report GIT-GVU-09-02, Georgia Institute of Technology.

WINFIELD'S ETHICAL ROBOT

Winfield et al. 2014. Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In Advances in Autonomous Robotics Systems, LNCS 8717, 85–96.

UTILITY FUNCTIONS



OUR ETHICAL REASONING IN UNFORESEEN CIRCUMSTANCES

Dennis et al. 2016. Formal Verification of Ethical Choices in Autonomous Systems. Robotics and Autonomous Systems, 77, 1-14.

ETHICAL POLICIES

- We have a set of ethical concerns which we rank: killing is worse than stealing is worse than lying.
- A plan, P1, is worse than another, P2, if
  - P1 violates an ethical concern and P2 doesn’t
  - The worst concern violated by P2 and not by P1 is less serious than the worst concern violated by P1 and not P2
  - The worst concerns are equally bad, but P1 violates more concerns than P2 does

BRAKE FAIL ON LINE UP

ϕ_1 = do not damage own aircraft (1),

ϕ_2 = do not collide with airport hardware (2),

ϕ_3 = do not collide with people (3),

ϕ_4 = do not collide with manned aircraft (4).

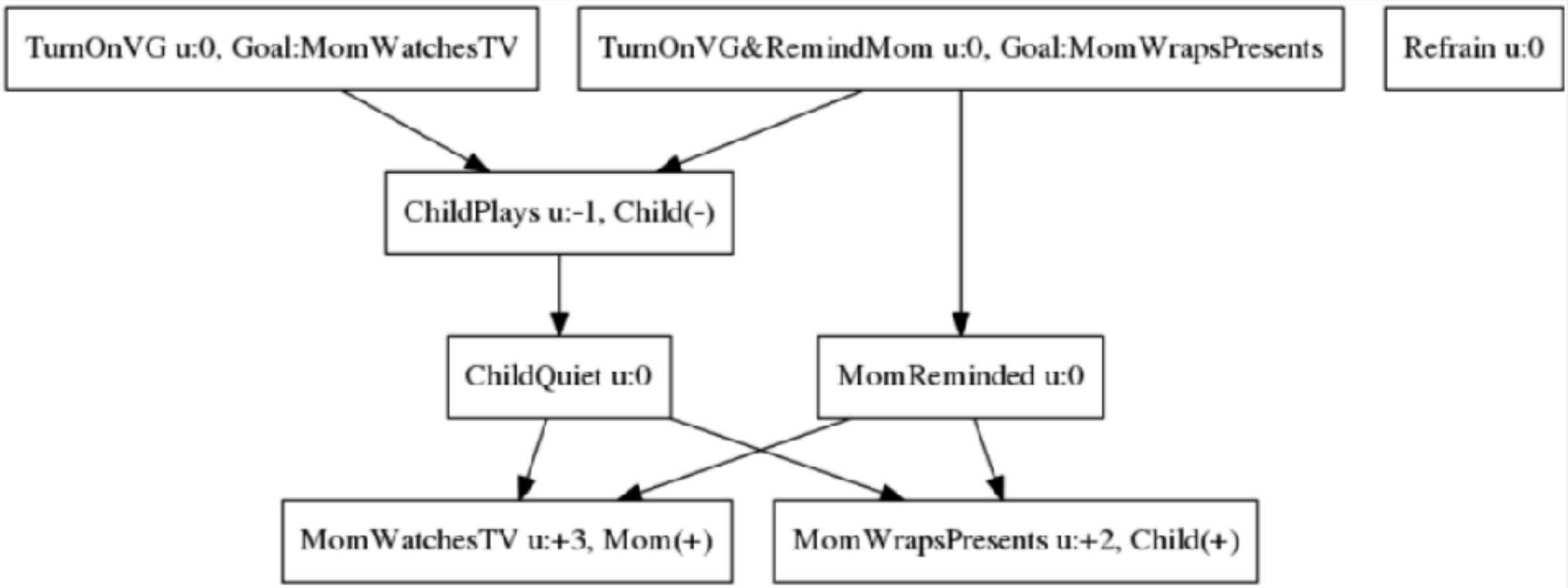
- Turn Left (damages the aircraft and airport hardware)
- Turn Right (damages the aircraft and risks colliding with people)
- Continue (risks collision with a manned aircraft)

AIRCRAFT TURNS LEFT

BENTZEN AND LINDNER'S HERA

COMPARING PHILOSOPHICAL ETHICAL SYSTEMS BY MODEL CHECKING

Immanuel: An Ethical Robot
The HERA Project: http://www.hera-project.com



A HERA MODEL

Bentzen et al. Moral Permissibility of Actions in Smart Homes. FLOC workshop on Robots, Morality and Trust through the Verification Lens, 2018

REFERENCES: GENETH

- Early Paper outlining approach:
  - Anderson et al, 2004. Toward Machine Ethics Proceedings of AAAI Workshop on Agent Organizations: Theory and Practice.
- Most recent instantiation and description:
  - Anderson et al, 2016. A Value Driven Agent: Instantiation of a Case-Support Principle-Based Behaviour Paradigm. Workshop in AI, Ethics and Society.
- Website: http://uhaweb.hartford.edu/anderson/Site/GenEth.html

REFERENCES: DCEC

- Description of Deontic Cognitive Event Calculus
  - Bringsjord & Govindarajulu (2013), Toward a Modern Geography of Minds, Machines, and Math, in V. C. Mueller, ed., 'Philosophy and Theory of Artificial Intelligence', Vol. 5 of Studies in Applied Philosophy, Epistemology and Rational Ethics, Springer, New York, NY, pp. 151–165.
- Akratic Robot:
  - Bringsford et al 2014. Akratic Robots and the Computational Logic Thereof. Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology.
- Website: https://rair.cogsci.rpi.edu/projects/muri/

REFERENCES: ETHICAL GOVERNOR

- Original Description:
  - Arkin et al. 2009. An Ethical Governor for Constraining Lethal Action in an Autonomous System. Technical Report GIT-GVU-09-02, Georgia Institute of Technology.
- Recent application in Healthcare:
  - Shim and Arkin (2015). An Intervening Ethical Governor for a Robot Mediator in Patient-Caregiver Relationships.A World with Robots: International Conference on Robot Ethics: ICRE 2015

REFERENCES: WINFIELD'S ROBOTS

- Original Paper:
  - Winfield et al. 2014. Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In Advances in Autonomous Robotics Systems, LNCS 8717, 85–96.
- More Recent version:
  - Vanderelst and Winfield (2017). An Architecture for Ethical Robots inspired by the Simulation Theory of Cognition. Cognitive Systems Research.

REFERENCES: ETHICAL CHOICES IN UNFORESEEN CIRCUMSTANCES

- Dennis et al. 2016. Formal Verification of Ethical Choices in Autonomous Systems. Robotics and Autonomous Systems, 77, 1-14.

REFERENCES: HERA

- Lindner, F.; Bentzen, M. M.. 2018. A Formalization of Kant's Second Formulation of the Categorical Imperative. Accepted for publication in *The proceedings of the 14th International Conference on Deontic Logic and Normative Systems (DEON 2018)*.
- Lindner, F.; Bentzen, M. M.; and Nebel, B. 2017. The HERA approach to morally competent robots. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017, 6991–6997.
- Bentzen et al. Moral Permissibility of Actions in Smart Homes. FLOC workshop on Robots, Morality and Trust through the Verification Lens, 2018