Testing Ethical Decision-Making in Autonomous Systems

Mohammad Mousavi
mohammad.mousavi@kcl.ac.uk
Jan65 Fest!
03 March 2023

Based on joint work with: Michael Akintunde, Martim Brandao, Gunel Jahangirova, Hector Menendez, and Jie Zhang

For Jan Peleska,
Who has a passion for going through all scenarios,
While standing on firm grounds….

Happy Birthday Jan and
Many happy scenarios!

Prelude:
Trust in Autonomous systems

MISSION STATEMENT

We apply **rigorous analysis** techniques
to ensure **safety** and
to establish **trust** in **autonomous systems**.

Featured In:

BBC RADIO 4

ALJAZEERA

THE ROYAL SOCIETY

Discovery EDUCATION

LONDON MATHEMATICAL SOCIETY EST. 1865

Sponsors:

UKRI Engineering and Physical Sciences Research Council

UKRI Science and Technology Facilities Council

UKRI Trustworthy Autonomous Systems Hub

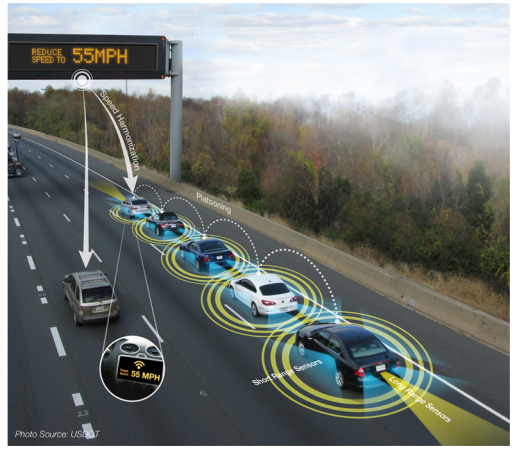
Trust: Definition and Facets

- **User's belief** that a system's performance is **helpful and safe**, particularly in **uncertain and critical situations**

- Major contributors:
  - Technical **transparency** (design intentions)
  - **Explainability**
  - **Safety** and **Security**
  - **User experience**

Connected Platoons

- Main idea: **rigorous safety analysis** of design space for autonomous functions
- Approach:
  - **Conformance testing**: comparing the behaviour of an "ideal platoon" with a parameterised implementation
  - **Genetic algorithms** and **search-based** techniques
    - **Maximising risk**
      pushing the system towards failure
  - Analysing the outcomes and finding **optimal parameters**
- Application:
  - Increase **trust** by improving the **safety** and **performance** of **basic awareness protocols** in connected autonomous functions



[Araujo, Hoenselaar, MRM and Vinel. PIRMC 2020]

Particle Emission Tests

- Main Idea: design effective **particle emission tests** using a combination of **machine learning** and conformance **testing**

- Approach:
  - Use **real data** (from NOx sensors) to **learn** a model of **vehicle's behavior**
  - Run **search-based** techniques to **maximise emission** under given **test constraints**

- Applications:
  - Detecting cheaters: **doping tests**, e.g., for diesel cars
  - Designing **environment-friendly** profiles for **autonomous vehicles**

- ***Trust*** *through* ***transparency***

[Dimitrova, Gazda, MRM, Biewer, Hermanns, FORTE 2020]
[idem, Fries, Heinze, LMCS 2022]

Adaptive Model Learning

- Main idea: **learn succinct models** that summarise **spatial** and **temporal evolution** of systems
- Approach:
  - Carefully observing **redundant** and **deprecated** queries in automata learning
  - Summarising models learned from **various products**; making the process efficient for **small sample sizes**

- Application:
  - Increase **trust** by providing **models** explaining **system evolution**

[Damasceno, MRM and Simão. EMSE 2021]
[Tavassoli et al. SPLC 2022] (Best paper)
[Labbaf et al. FOSSACS 2023]

KASPAR EXPLAINS

- Main Idea: how **causal explanation** can influence stakeholders' trust
- Approach:
  - Analysing **interaction videos**
  - **Formalising interactions** in behavioural models
  - Generating **causal explanations** for various stakeholder

- Applications:
  - Kaspar: an educational **robot** for **autistic children**

- ***Trust*** *through* ***explanation***

KASPAR

University of Hertfordshire UH

COMPUSULT

UKRI Trustworthy Autonomous Systems Hub

http://bit.ly/KasparExplains

[Araujo et al. ICSR 2022]

[Sarda Gou et al. RO-MAN 2022]

https://verifiability.org

Zooming In: Testing Ethics

Ongoing joint work with: Michael Akintunde, Martim Brandao, Gunel Jahangirova, Hector Menendez, and Jie Zhang

Testing Ethics: Why?

- Machines making ethically-charged decisions:
  - Environmental impact of vehicle control
  - Fairness implications of credit scoring and financial decisions

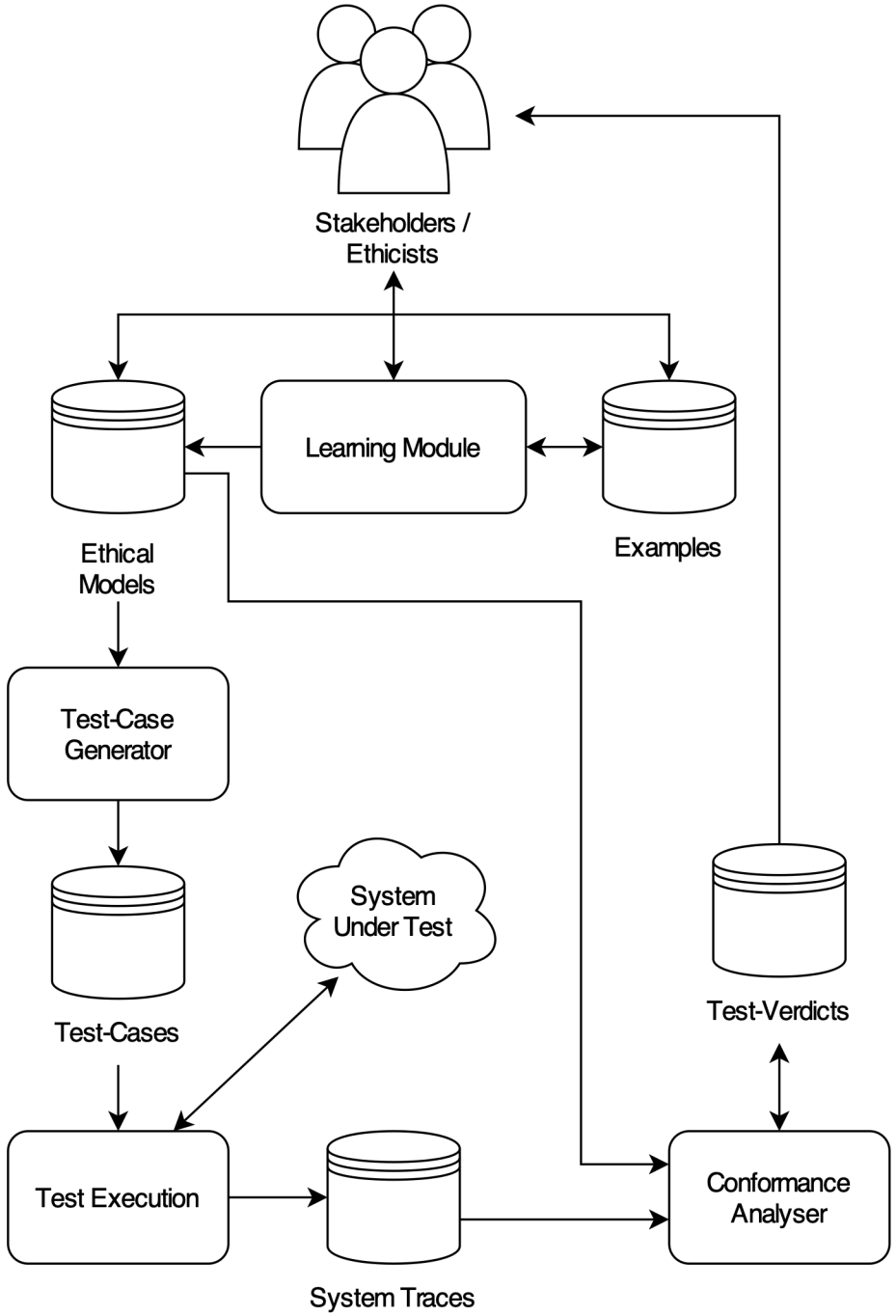
- Need for a discipline of testing cases and oracles improve
  - Transparency and balance conflicting interests of stakeholders
  - Understanding ethics

"AI makes Philosophy honest."
– D. Dennett Computers as Prostheses for Imagination

Testing Ethics: Challenges

- Generating effective scenarios
- Different meta-ethical frameworks
- Ethical oracles
- Diversity and stakeholder engagement

Architecture



Testing Ethics: Running Example

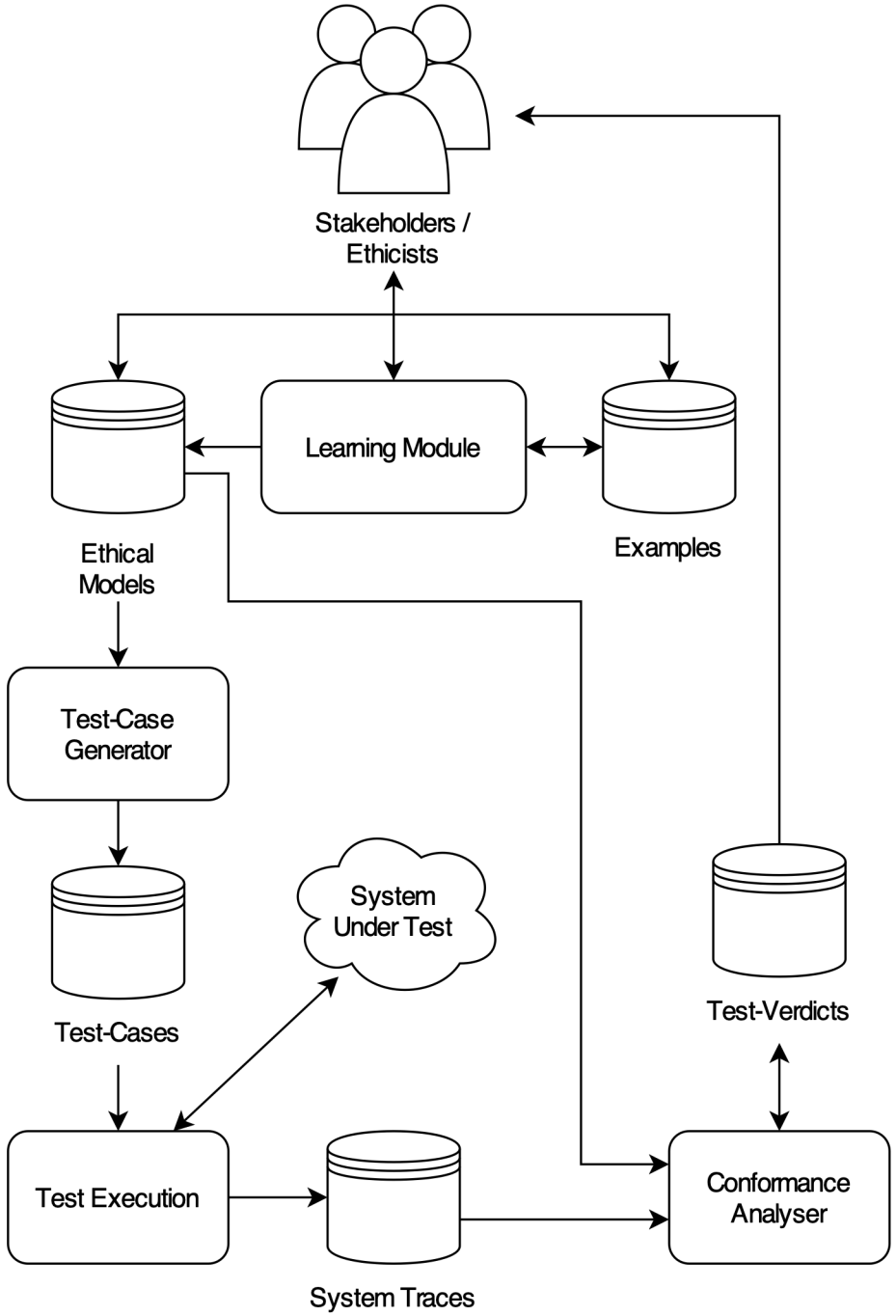
Copyright: C. Junker, Flickr, CC2.0

Scenario Generation Ethics: Challenges

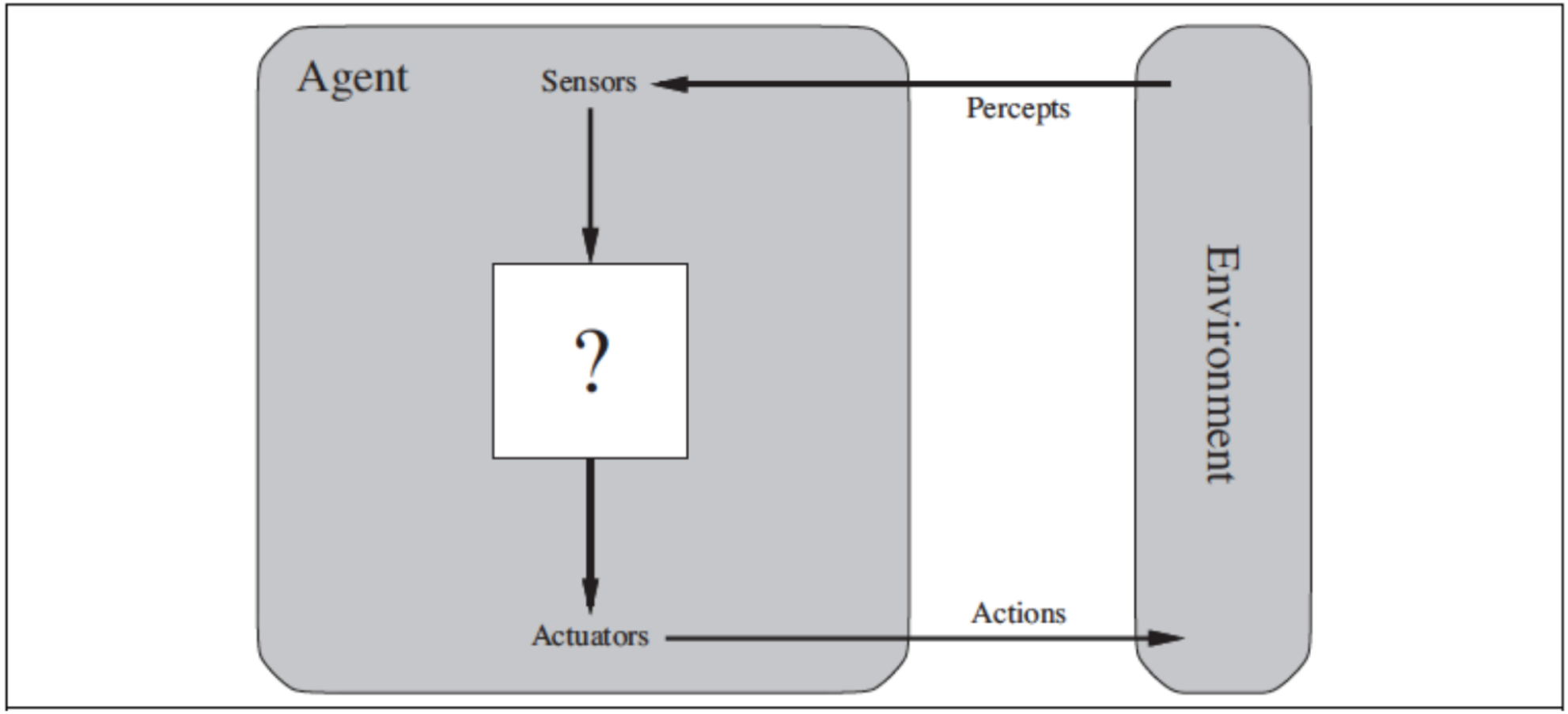
- Objectives:
  - Efficient / effective fault detection
  - Useful vehicles for user engagement
- Diversity and coverage

- Scenarios that:
  - conformance to various parts of the highway code,
  - choices between the code and giving way emergency vehicles
  - Covering various combinations of ethical objectives / regulations

Architecture



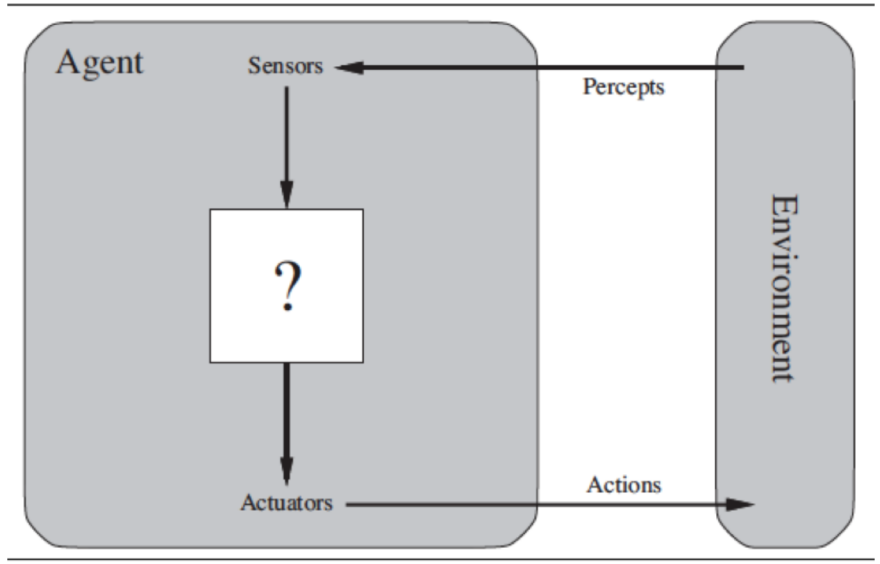
Test Oracles: Metaphor



From: Russel and Norvig, Artificial Intelligence: A Modern Approach. p. 35 , Prentice Hall.

Testing Ethics: How?

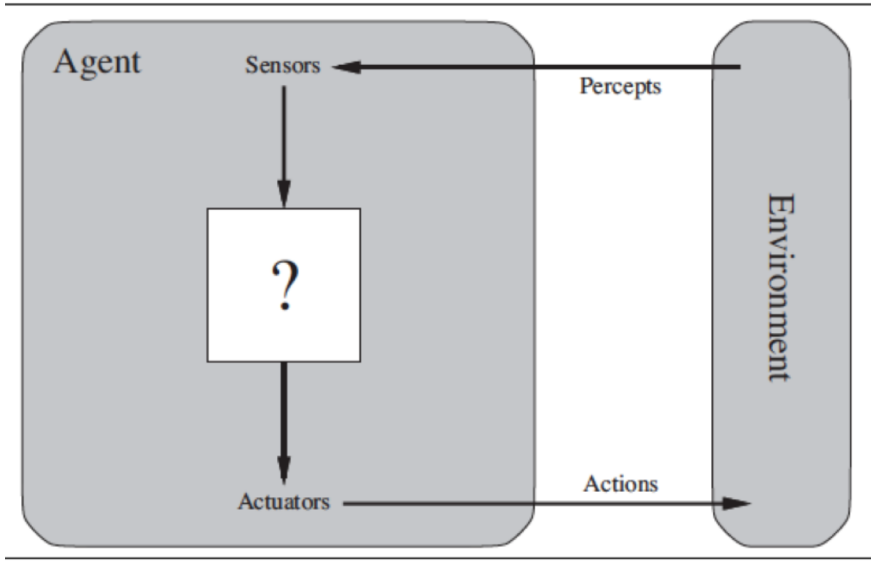
- Meta-ethics:
  - Deontologicism: acts are inherently good or evil
  - Consequentialism (Utilitarianism):acts are good or evil because of (depending on) their consequences
  - • Virtue ethics: an action is good or evil if it it fits for (improves) a virtuous person
- Many variants, taking various environmental conditions into account:
  - Prima facie duty



Testing Ethics: How?

- Oracle for a deontological theory
  - **Objective** $Obj: Action \rightarrow Value$
  - **Test input** $\alpha: Percept^*$
  - **Test output** $a\ after\ \alpha$: $Action$. (last observed action)
  - **Verdict** $|Obj(a\ after\ \alpha) - Obj\ (ideal\ after\ \alpha)| \leq \varepsilon$
- Problems:
  - Values of actions:

    Partial orders among actions
  - Optimal α : search-based testing
  - Correct $ideal\ after\ \alpha$: • Surrogate models
  - Threshold ε : domain knowledge



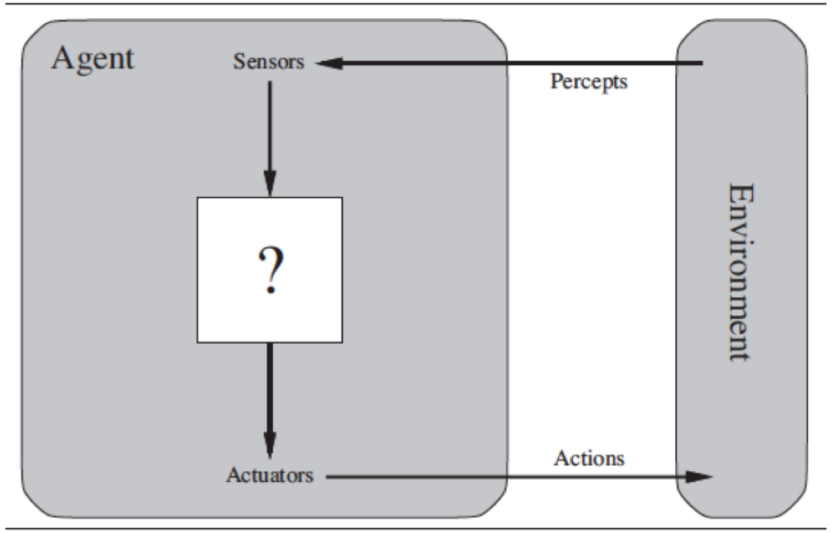
[Ganascia, Ethical System Formalization using Non-Monotonic Logics, 2007]

Testing Ethics: How?

- an agent shall respect human lives,
  obj(kill) = -1
- an agent shall give way to ambulances,
  obj(give_way) = +1 and
- an agent shall not damage other cars
  obj(damage) = -1

Testing Ethics: How?

- Oracle for act-utilitarianism:
  - **Objective** $Obj: Action \times Env \rightarrow Value$
  - **Test input** $(\alpha, env): Percept^* \times Env$
  - **Test output** $a\ after\ \alpha: Action$. (last observed action)
  - **Verdict** $|Obj(a\ after\ \alpha, env) - Obj\ (ideal\ after\ \alpha, env)| \leq \varepsilon$
- Problems:
  - Similar problems as to deontological ethics
  - Richer value models
  - Conflicts with deontological theory

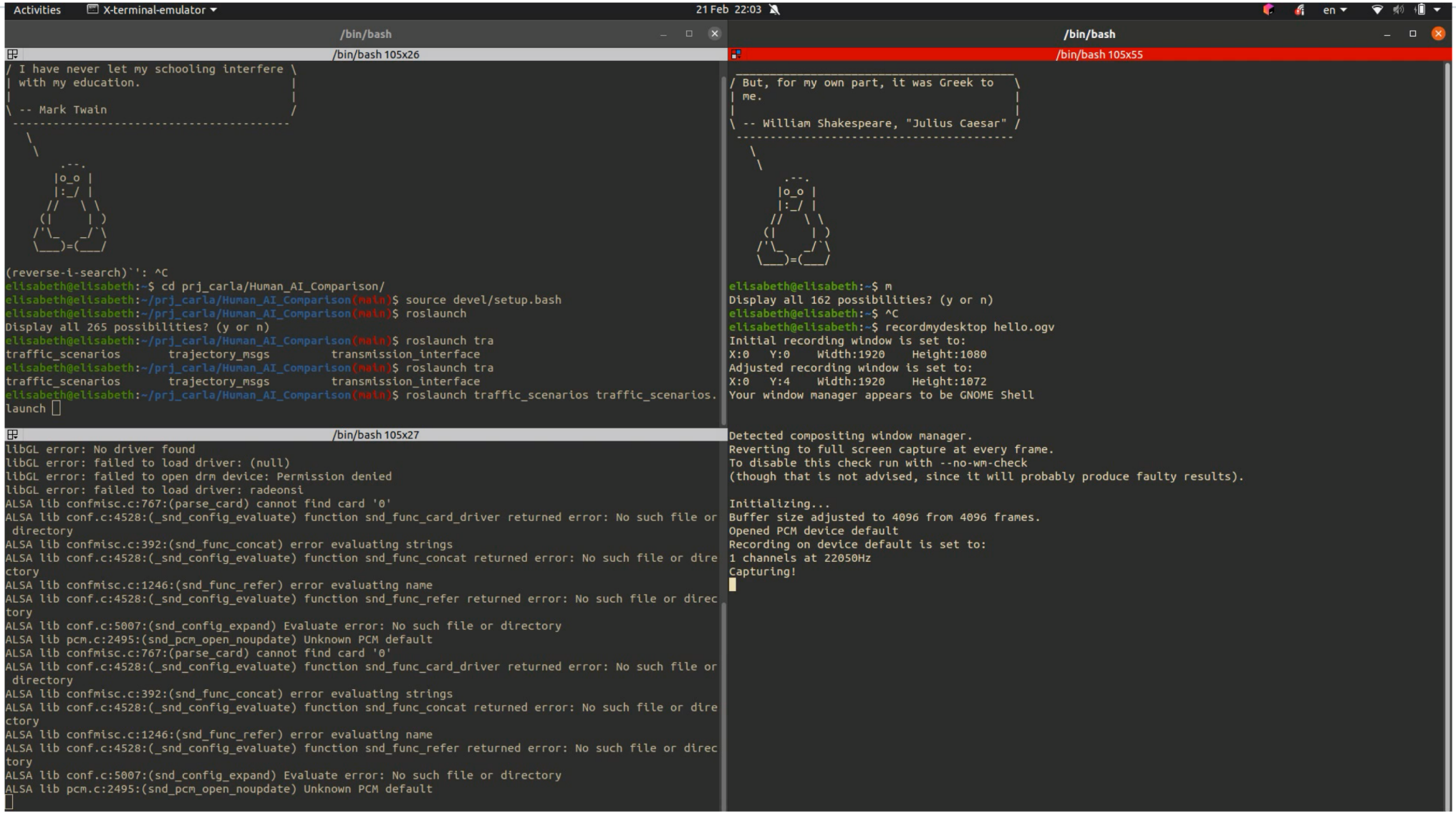


[Shea-Blymyer and Abbas. Algorithmic Ethics: Formalization and Verification of Autonomous Vehicle Obligations. HSCC 2020 and Arxiv.]

Testing Ethics: How?

- an agent shall respect human lives at all times,
  $\forall env$. obj(kill, env) = -100
- an agent shall give way to ambulances if it does not involve a damage to other cars,
  $\forall env.\ giveway \bowtie damage$
  $\Longrightarrow obj(giveway, env) = +10$

Testing Ethics: How?



Elisabeth Halvorsen, Bachelor Thesis, KCL.

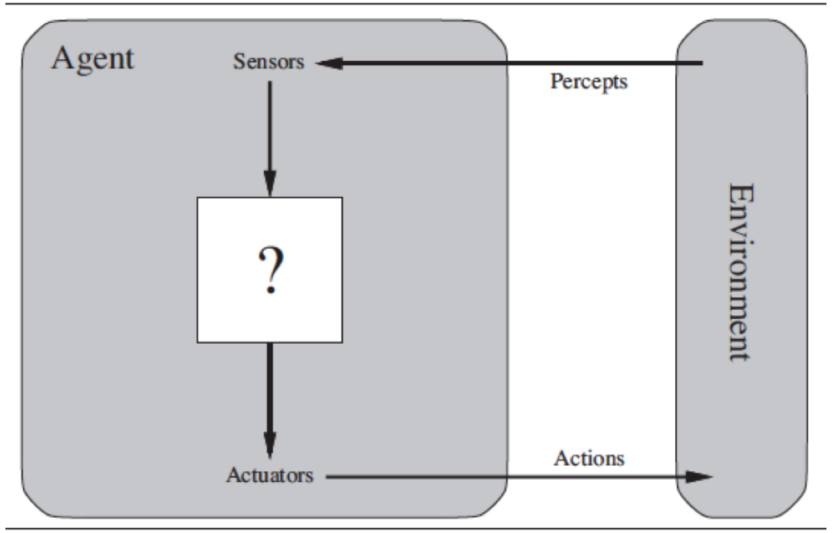
Testing Ethics: How?

- Oracle for virtue-ethics:
- Model of a moral exemplar S (robotic saint!) • Conformance btw. agent A and exemplar:
- distance of their
  - states
  - output sequences

  under the same input sequence.
- $\forall \alpha: (Percept^* \times Action)^*. \alpha \in beh(agent)$

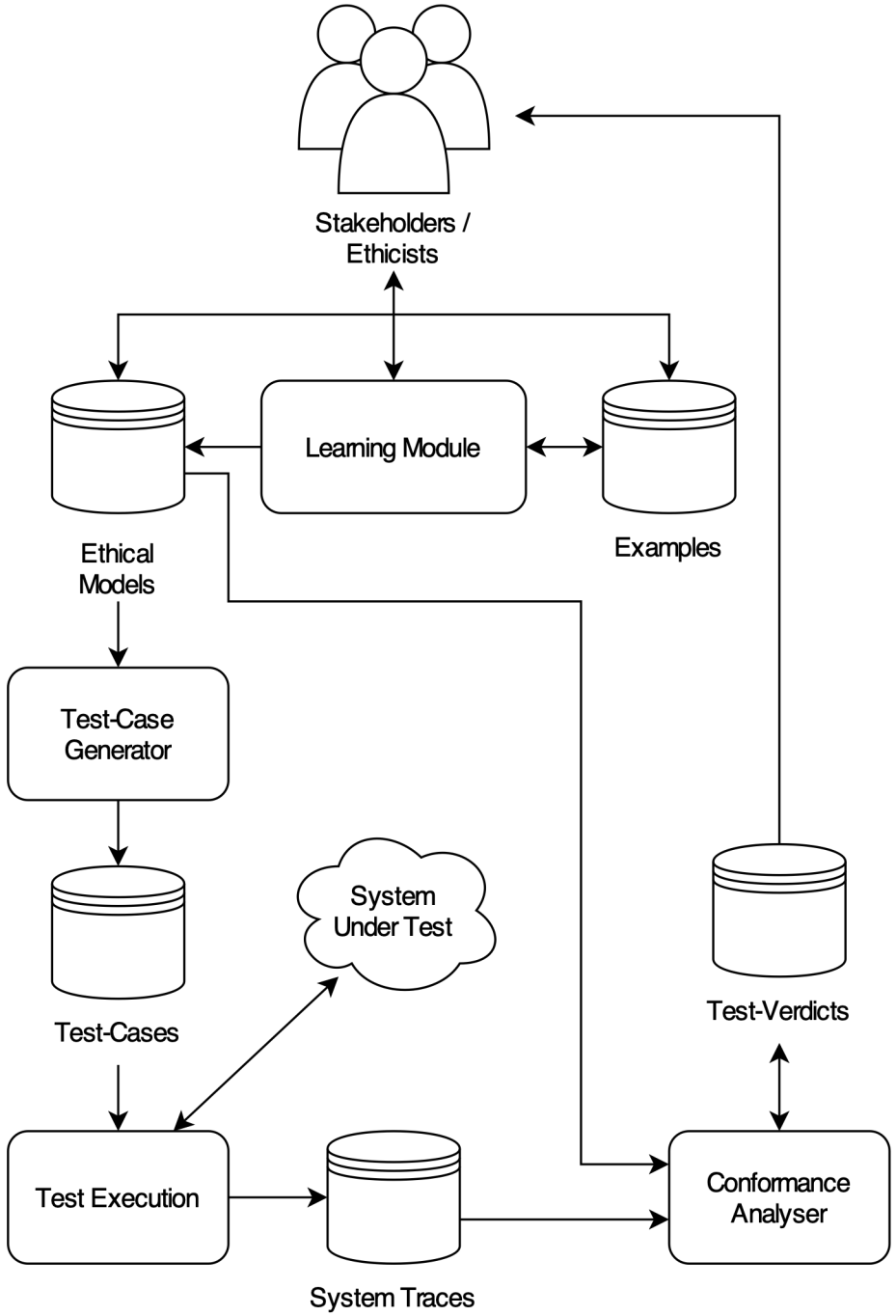
  $\forall \beta: (Percept^* \times Action)^*. \beta \in beh(saint).$

  $dist(\alpha, \beta) \leq \epsilon$



[Govindarajulu et al. Toward the Engineering of Virtuous Machines, AIES 2019]

Architecture



Thank you

Mohammad Mousavi
mohammad.mousavi@kcl.ac.uk