ETH zürich

Liability, ethics, and culture-aware behavior specification using rulebooks

Andrea Censi

About me

- M.Eng. @ **Sapienza**, Control systems and robotics
- Ph.D. @ **Caltech**, 2012, Control and Dynamical Systems
- P.I. @ **MIT**, 2013-16
- System Architect @ **nuTonomy**

currently:

- Senior Researcher @ **ETH Zürich**
- Director of Research @ **Aptiv Autonomous Mobility**
- President @ **Duckietown Foundation**

The opinions described here are the speaker’s own, and not necessarily representative of any employer’s position.
The functionality described is not necessarily representative of current and future products by Aptiv and its partners.
The scenarios discussed are simplified for the purposes of exposition and do not fully capture internal safety processes.

About Aptiv

- **Aptiv**
  - Formerly known as “Delphi”.
  - Tier 1 supplier
  - >150k employees

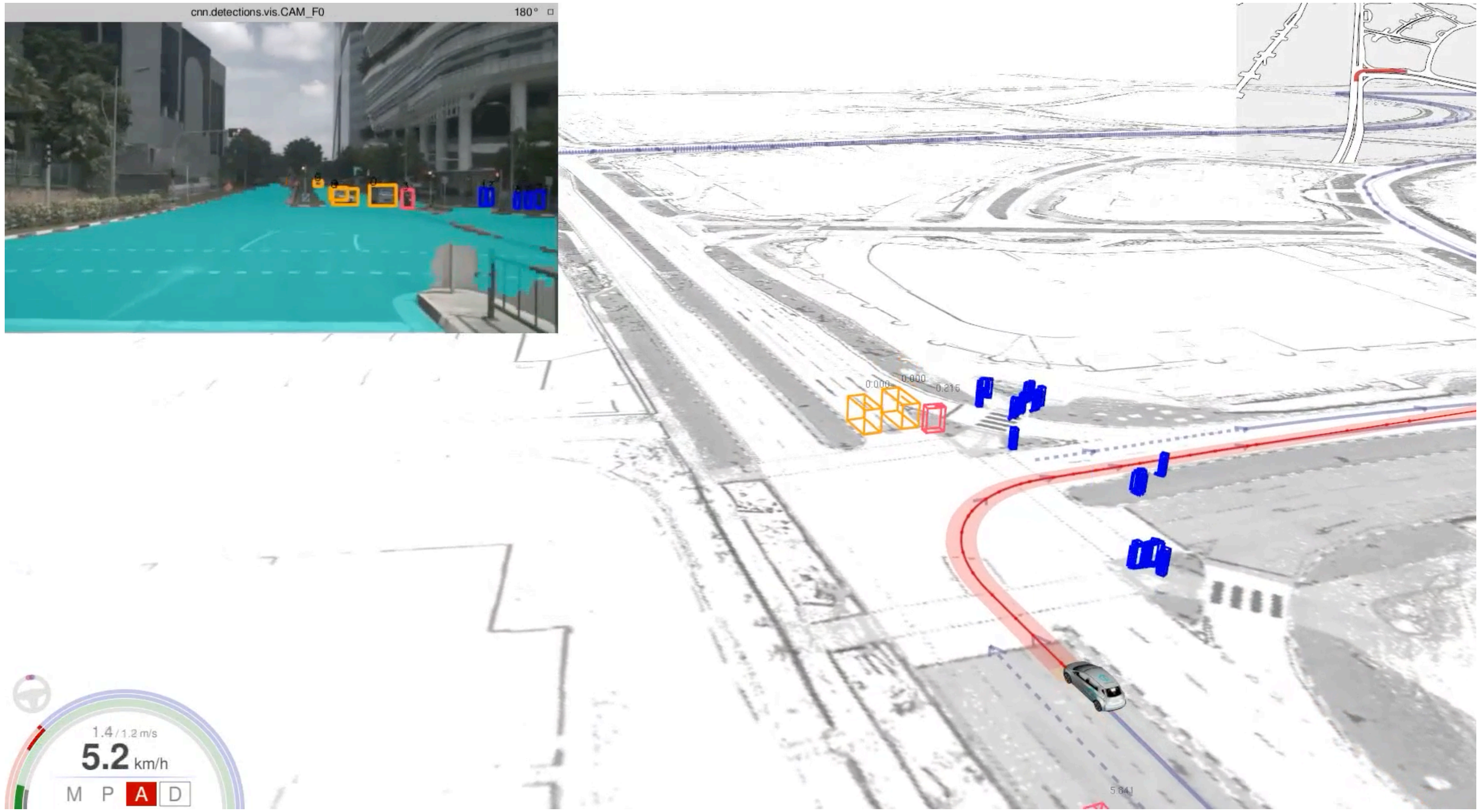
- **Aptiv AM (Autonomous Mobility)**
  - ~800 people
  - Development in Boston, Pittsburgh, Singapore, Santa Monica.
  - Public deployment in Las Vegas.
    - 30,000 rides, 1M+ miles
    - 4.92 stars ★★★★★

- **Aptiv’s engagement with the research community**
  - *nuScenes* (`nuscenes.org`), richest sensor dataset available for AVs.
  - *AI Driving Olympics* (NeurIPS 2018, ICRA 2019 announced this week)

Overview

1. **Making AVs** is difficult
2. **Decision making** is difficult
3. **Behavior specification** is difficult
4. The **rulebooks** approach to behavior specification
5. Open issues

nuTonomy product (early 2017)



Why is making an AV difficult?

- In addition to what you can guess:

- Because component metrics are not predictive of system metrics.
- Because abstractions and interfaces are “leaky”.
- Because it is an “open system” that you cannot fully know.
- Because no single modeling technique is sufficient to capture everything important.
- Because no single algorithm covers all the operating domain.
- Because hacks compound.
- Because the task is complex and nuanced.

- **At industrial scale, broader systems issues start to dominate on narrower algorithmic issues.**

A new kind of engineering

- **Engineering is inching closer to the natural sciences:**
  We create things that we do not fully understand,
  then investigate our creations.

Perception is conceptually simple; decision making is not

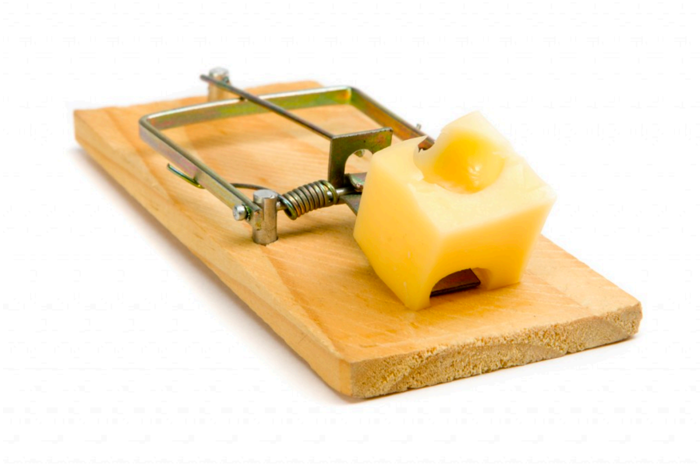
Perception is simple		Decision making is complex
sensor data, sensor models, priors, Bayes Filter	**formalization**	interactive/closed loop. partially observable world, other agents with unknown intent, partially non-cooperative
ground truth, well-understood metrics	**specification**	no ground truth, conflicting requirements
from the comfort of your desk	**development**	need to hit the road, hard to test in isolation

Behavior requirements are numerous, vague, and conflicting

- **Function** (Pick up, drop off, etc.)
- **Compliance to traffic rules**
  - extensive & diverse
  - written to be read by humans
  - applicable to human drivers
- **Safety**
- **Liability**
- **Courtesy**
- **Comfort**
- **Culture**
- **Ethics**

"Does your car have any idea why my car pulled it over?"

Traps to avoid for AV behavior specification



✗ **Hard constraints**

- “Infeasibility” is not a thing for embodied intelligence
- Other actors prevent most guarantees.

✗ **Case analysis, finite state machines, …**

“IF statements kill people”

✗ **Just relax, man**

$$J = \alpha J_1 + \beta J_2 + \gamma J_3 + \dots$$

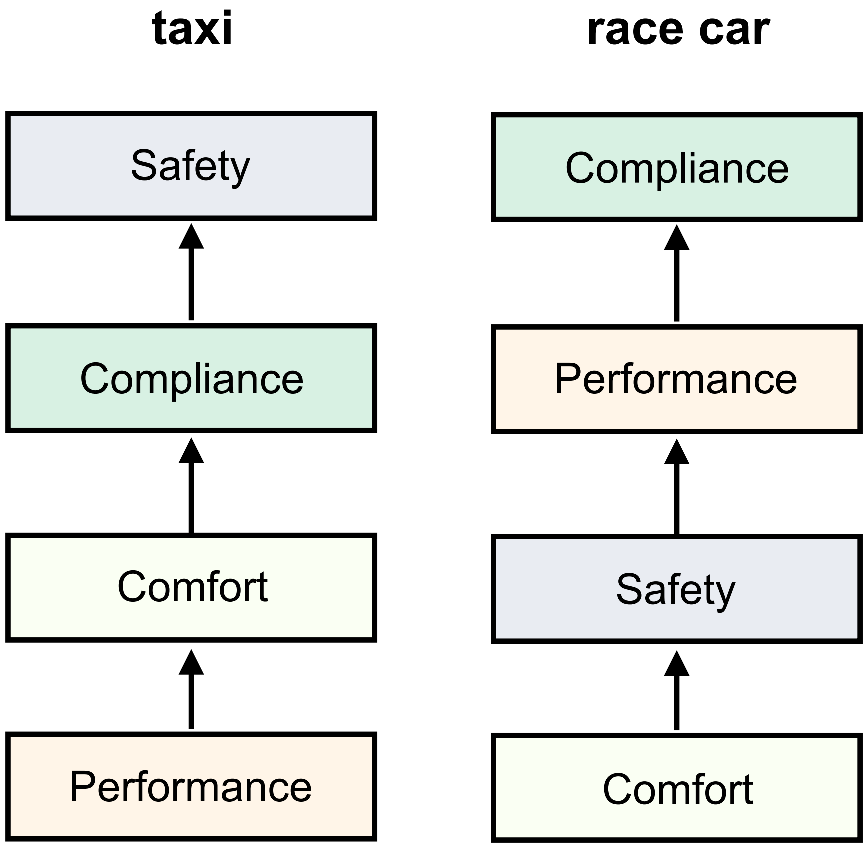
- Hard to re-tune; prone to overfitting.
- Lack of transparency.

Minimum violation planning

- Assume that constraints will be violated;
  find the alternative that least violates them.

  1. Define **rules** as a **total order** over realizations;
  2. **Order rules** according to priority;
  3. Obtain a **lexicographic order** for realizations.

- Allows **modular definition** of behavior.
- **Easy to predict** what the car will do.
- **Easy to understand** why the car did something.
- Introducing “tolerances” improves expressivity and leads to a lexicographic semi-order.

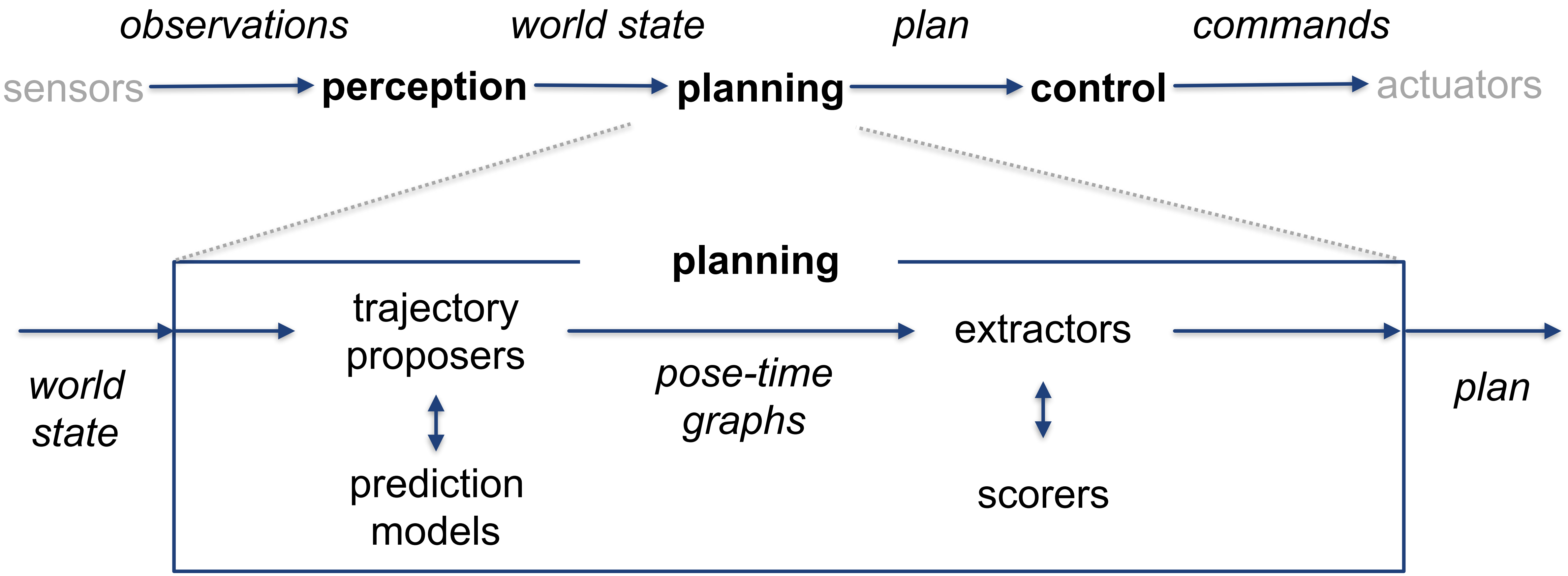


Planning using unbridled creativity and good taste

"The way to get good ideas is to get lots of ideas, and throw the bad ones away." — Linus Pauling

"creativity" + **"good taste"**

Planning using unbridled creativity and good taste



“creativity”

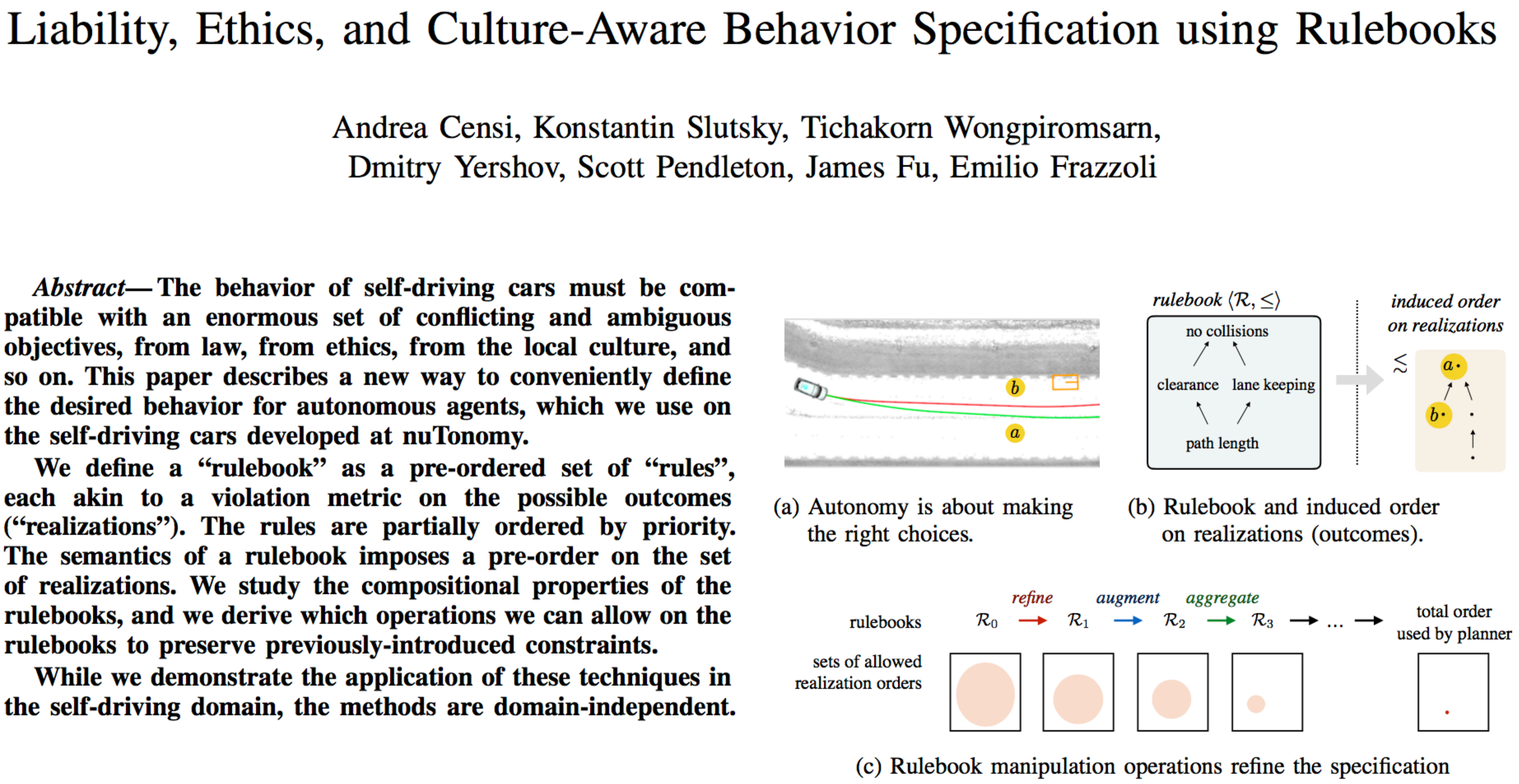
“good taste”

Good specifications specify little

- For behavior specification, do we need to choose an exact ordering of hundreds of rules?

- In the “**rulebooks” formalization**:
  - We use coarser resolution using **rule groups;**
  - We use **pre-orders** instead of total orders.

To appear at ICRA 2019; preprint available on Arxiv



Liability, Ethics, and Culture-Aware Behavior Specification using Rulebooks

Andrea Censi, Konstantin Slutsky, Tichakorn Wongpiromsarn,
Dmitry Yershov, Scott Pendleton, James Fu, Emilio Frazzoli

***Abstract*—The behavior of self-driving cars must be compatible with an enormous set of conflicting and ambiguous objectives, from law, from ethics, from the local culture, and so on. This paper describes a new way to conveniently define the desired behavior for autonomous agents, which we use on the self-driving cars developed at nuTonomy.**

We define a "rulebook" as a pre-ordered set of "rules", each akin to a violation metric on the possible outcomes ("realizations"). The rules are partially ordered by priority. The semantics of a rulebook imposes a pre-order on the set of realizations. We study the compositional properties of the rulebooks, and we derive which operations we can allow on the rulebooks to preserve previously-introduced constraints.

While we demonstrate the application of these techniques in the self-driving domain, the methods are domain-independent.

(a) Autonomy is about making the right choices.

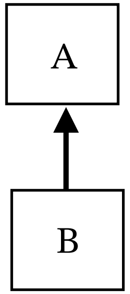
(b) Rulebook and induced order on realizations (outcomes).

(c) Rulebook manipulation operations refine the specification

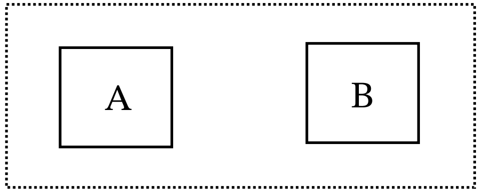
Rulebooks formalism (sketch)

- A ***rule*** is a total order on realizations.
- A **rulebook** is a pre-ordered set of *rules*.

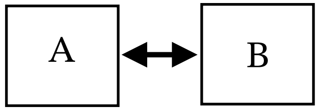
“Rule A is more important than rule B”



“The implementation can choose whether A or B is more important”

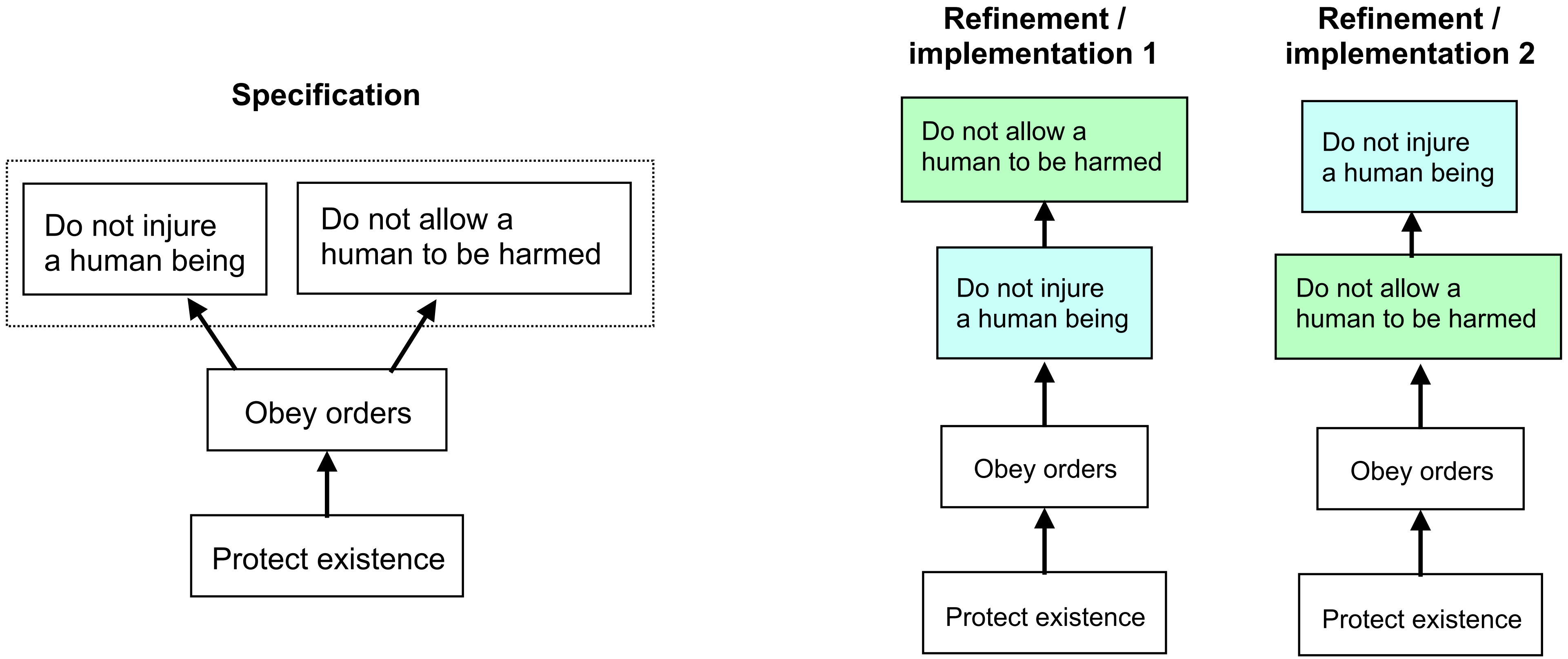


“Rule A and B must be at the same level”

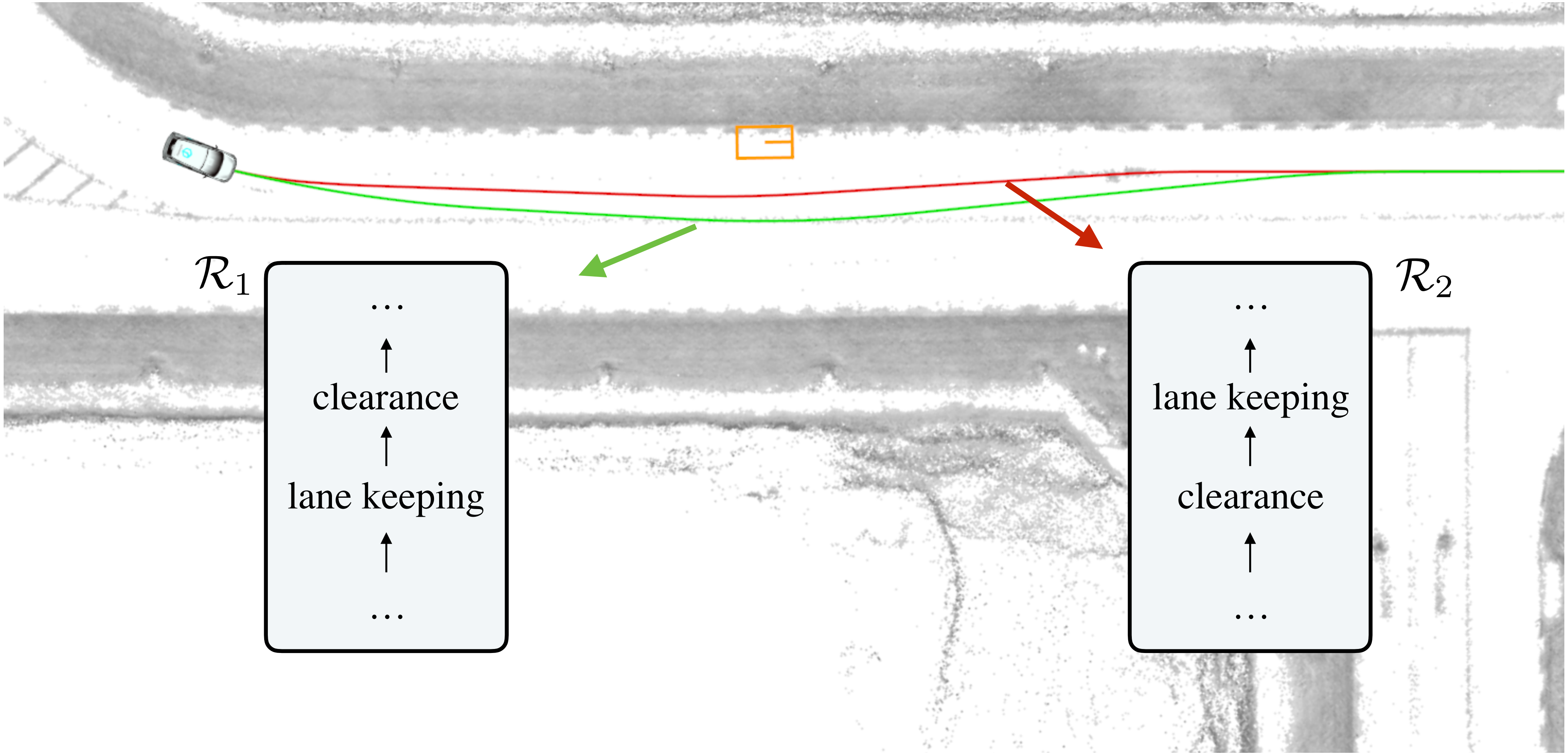


Obligatory Asimov example

1. *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
2. *A robot must obey orders given it by human beings, except where such orders would conflict with the First Law.*
3. *A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.*


Specification
Do not injure a human being
Do not allow a human to be harmed
Obey orders
Protect existence
Refinement / implementation 1
Do not allow a human to be harmed
Do not injure a human being
Obey orders
Protect existence
Refinement / implementation 2
Do not injure a human being
Do not allow a human to be harmed
Obey orders
Protect existence

User-friendly behavior shaping



$\mathcal{R}_1$

…

↑

clearance

↑

lane keeping

↑

…

$\mathcal{R}_2$

…

↑

lane keeping

↑

clearance

↑

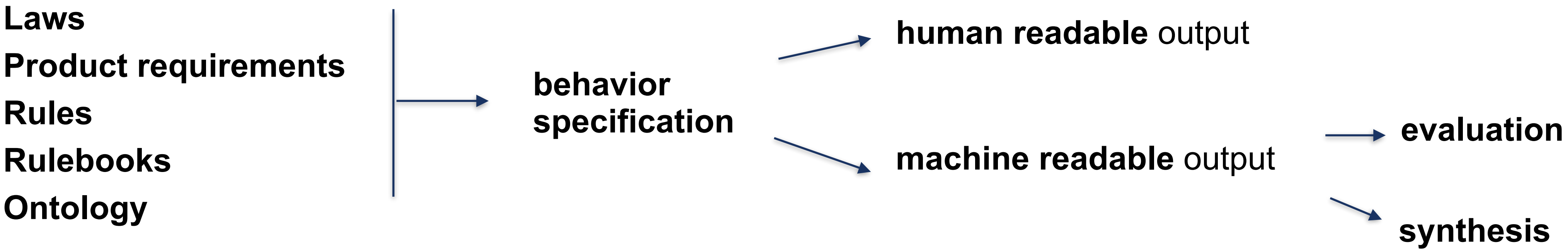
…

SLZ1476A

Rulebooks toolchain ensures traceability

- Ensure **traceability of behavior requirements** by **literated programming**:
  code (machine readable) and documentation (human readable)
  are close together and cross-referenced.

- **Domain-specific languages** describe rules and rulebooks.



Laws
Product requirements
Rules
Rulebooks
Ontology

behavior specification

human readable output

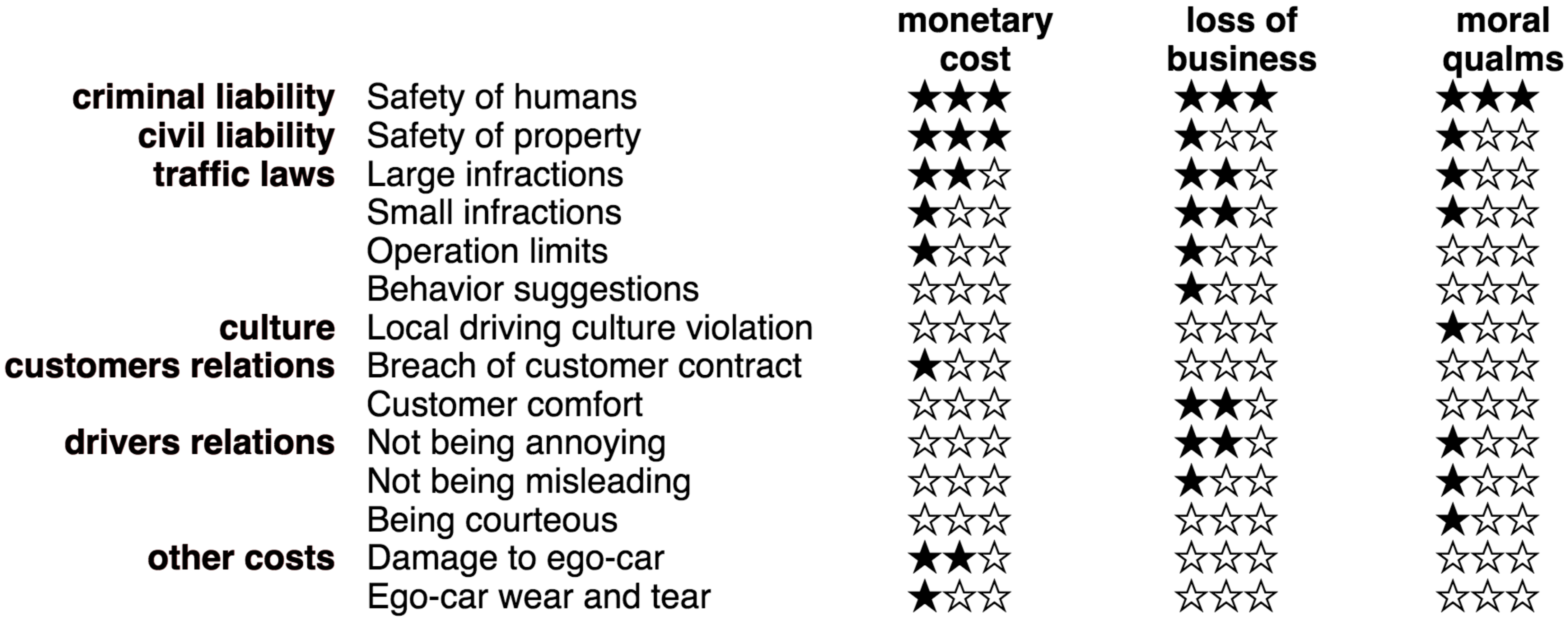
machine readable output

evaluation

synthesis

Defining and ordering rule groups

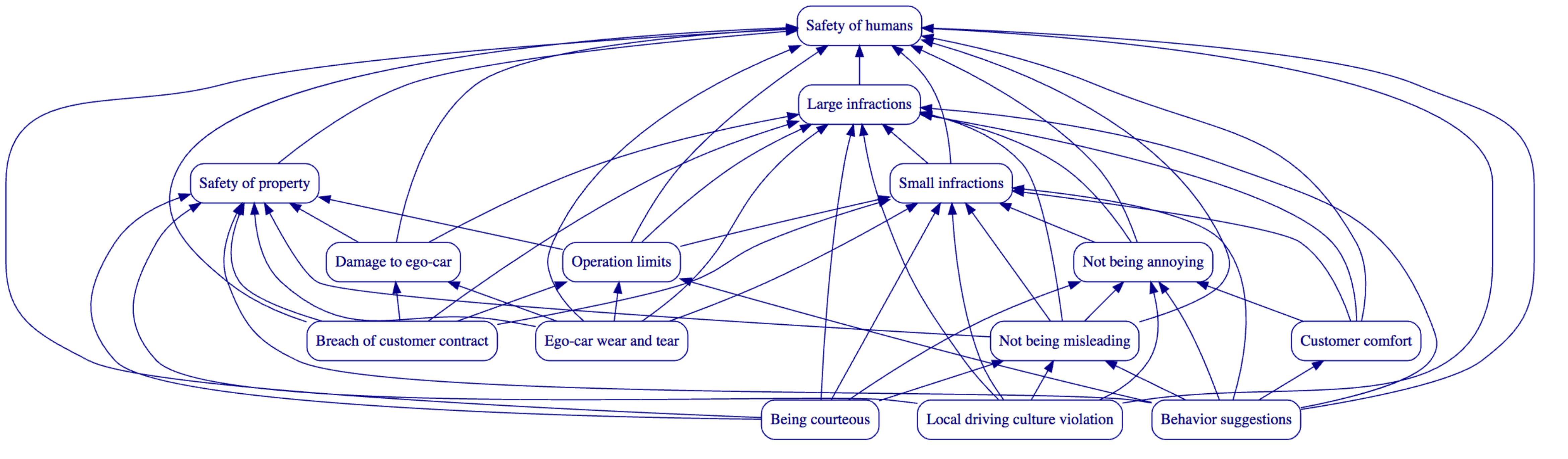
- Estimate: urban driving requires ~200 rules, ~20 rule groups.



		monetary cost	loss of business	moral qualms
criminal liability	Safety of humans	★★★	★★★	★★★
civil liability	Safety of property	★★★	★☆☆	★☆☆
traffic laws	Large infractions	★★☆	★★☆	★☆☆
	Small infractions	★☆☆	★★☆	★☆☆
	Operation limits	★☆☆	★☆☆	☆☆☆
	Behavior suggestions	☆☆☆	★☆☆	☆☆☆
culture	Local driving culture violation	☆☆☆	☆☆☆	★☆☆
customers relations	Breach of customer contract	★☆☆	☆☆☆	☆☆☆
	Customer comfort	☆☆☆	★★☆	☆☆☆
drivers relations	Not being annoying	☆☆☆	★★☆	★☆☆
	Not being misleading	☆☆☆	★☆☆	★☆☆
	Being courteous	☆☆☆	☆☆☆	★☆☆
other costs	Damage to ego-car	★★☆	☆☆☆	☆☆☆
	Ego-car wear and tear	★☆☆	☆☆☆	☆☆☆

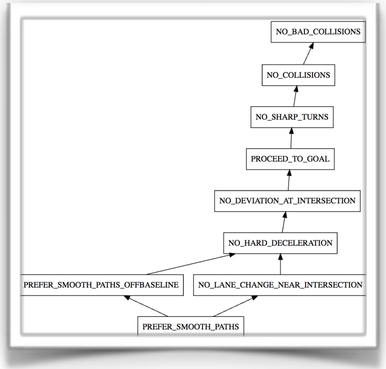
Defining and ordering rule groups

- Estimate: urban driving requires ~200 rules, ~20 rule groups.

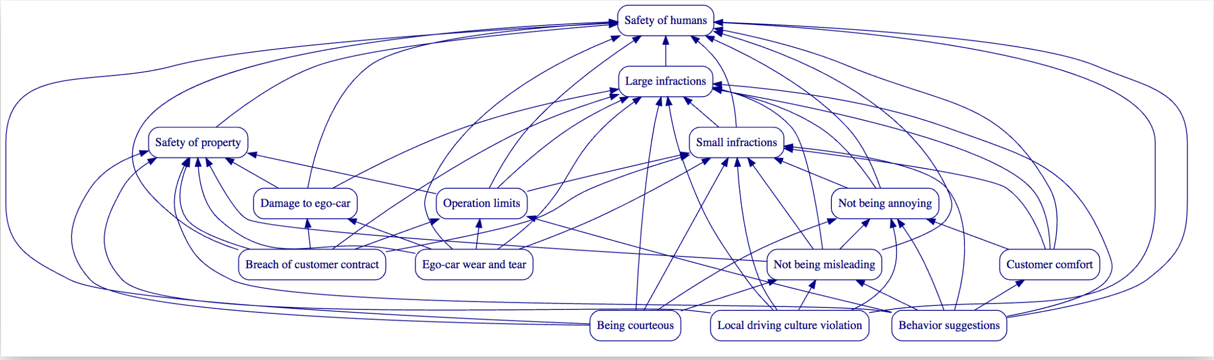


Example of automated analysis

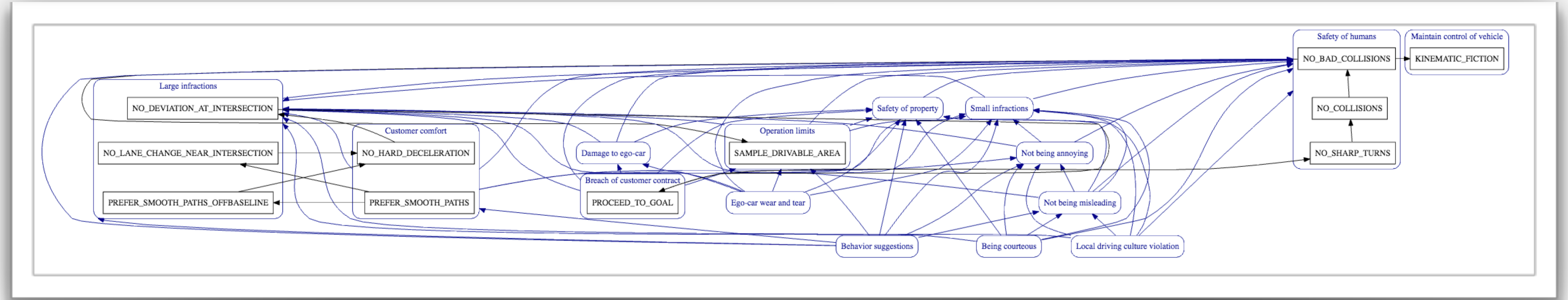
planner rulebook



rule group priorities



assign rules to groups

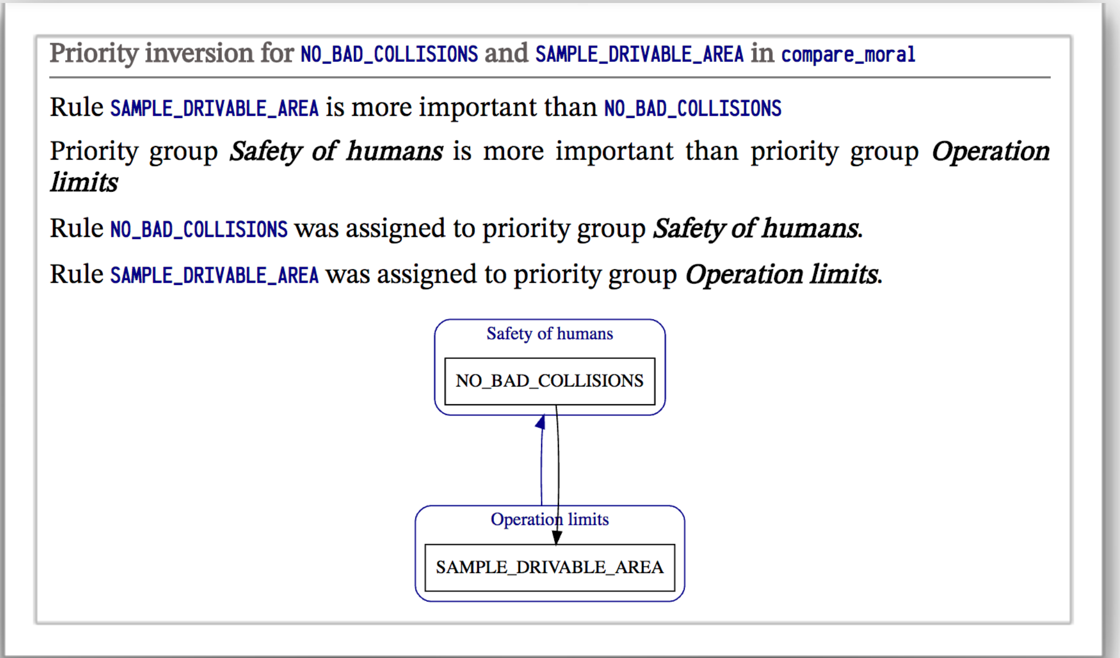


Example of automated analysis

- The result is tangled because there are

  cycles = priority inversions defects.

- These defects can be found automatically.



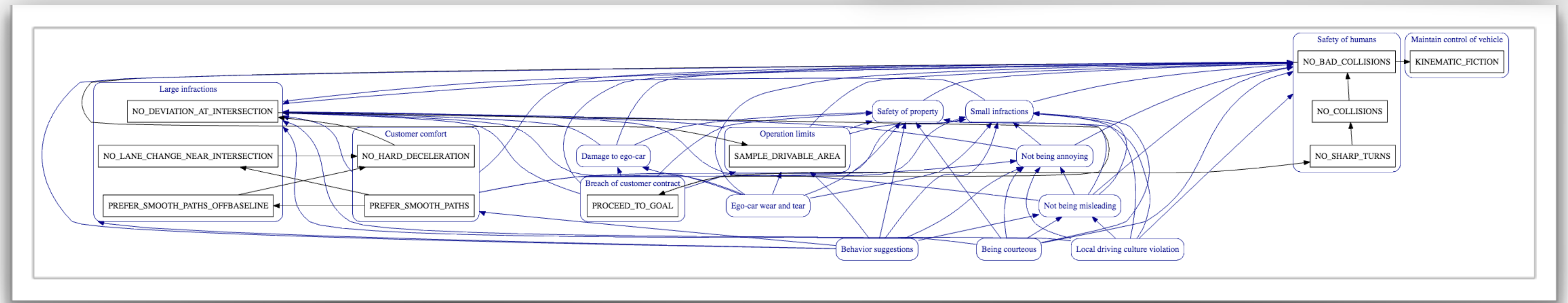
Priority inversion for NO_BAD_COLLISIONS and SAMPLE_DRIVABLE_AREA in compare_moral

Rule SAMPLE_DRIVABLE_AREA is more important than NO_BAD_COLLISIONS

Priority group *Safety of humans* is more important than priority group *Operation limits*

Rule NO_BAD_COLLISIONS was assigned to priority group *Safety of humans*.

Rule SAMPLE_DRIVABLE_AREA was assigned to priority group *Operation limits*.



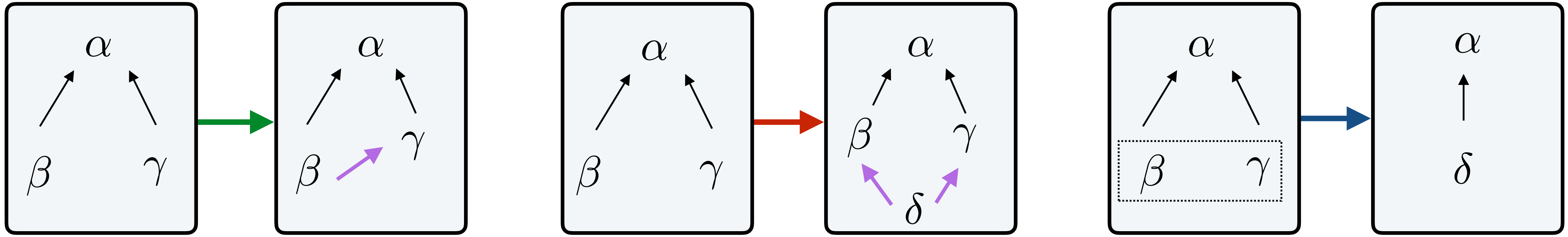
Customization

- Needs for customization:
  - **Local rules**;
  - Local **culture**;
  - **Function** flexibility (taxi *vs* truck);
  - **Customer** preferences.

- Wanted: a **compositional theory of behavior.**

Algebra of rulebooks enables compositionality

priority refinement *rule augmentation* *rule aggregation*

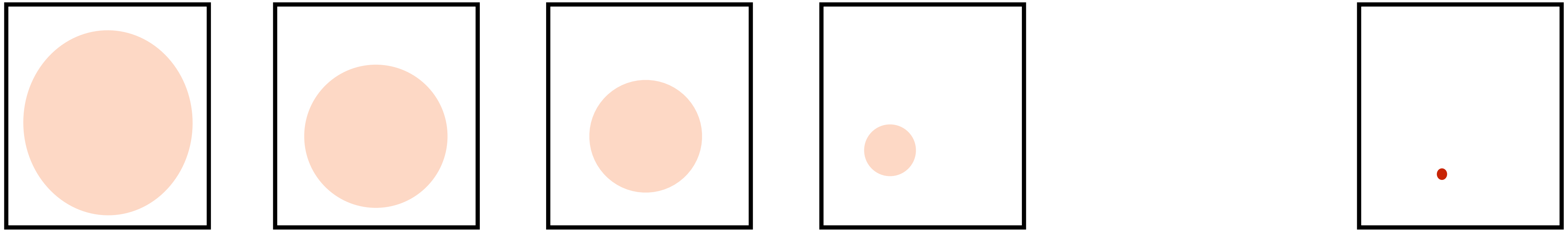


rulebooks

$$\mathcal{R}_0 \xrightarrow{\textit{refine}} \mathcal{R}_1 \xrightarrow{\textit{augment}} \mathcal{R}_2 \xrightarrow{\textit{aggregate}} \mathcal{R}_3 \rightarrow \dots \rightarrow$$

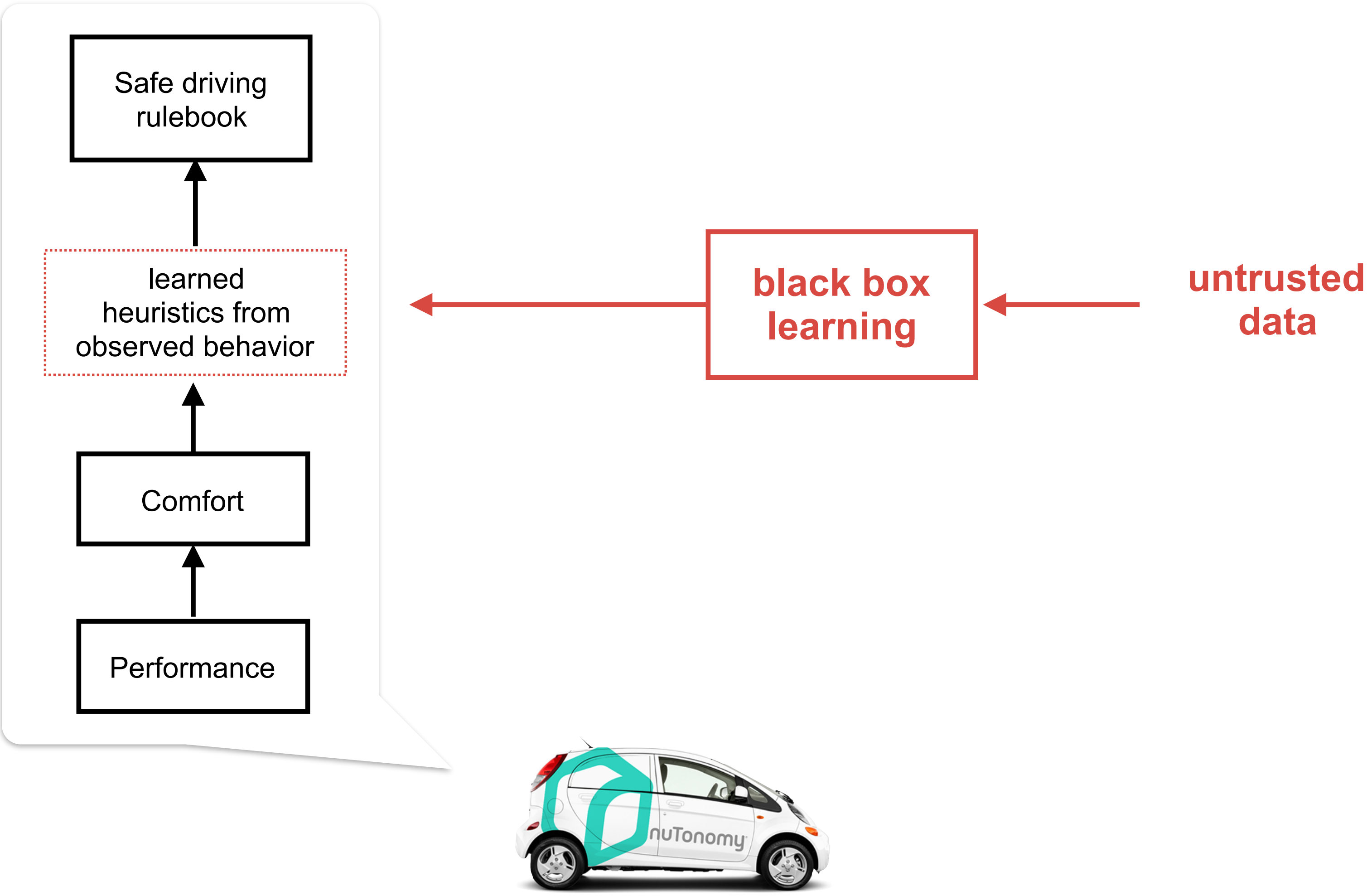
total order used by planner

sets of allowed realization orders



“Caging the learning”

- Rule priorities allow plug-and-play of untrusted heuristics without losing safety.



Safe driving rulebook

learned heuristics from observed behavior

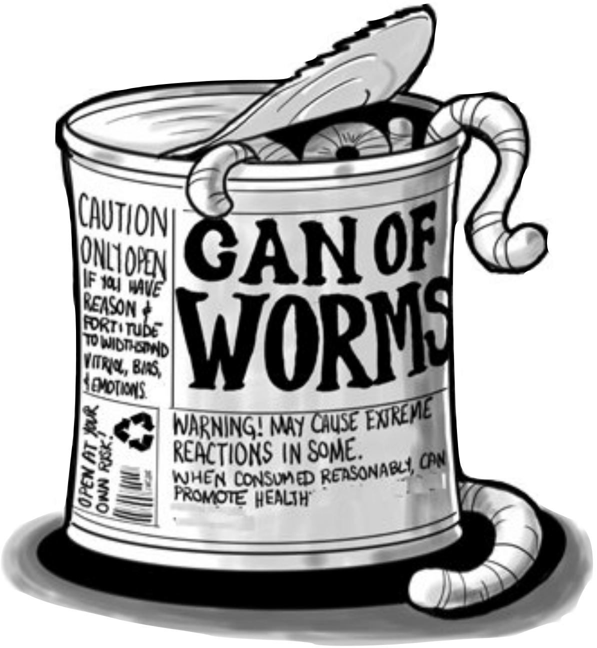
Comfort

Performance

black box learning

untrusted data

The ethical part

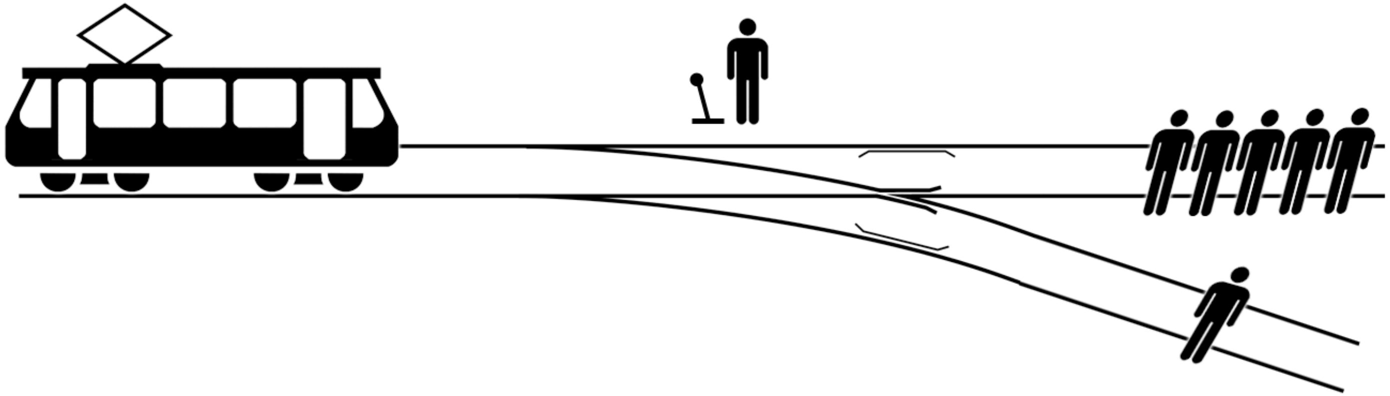

CAUTION ONLY OPEN IF YOU HAVE REASON & FORTITUDE TO WIDTHSTAND VITRIOL, BIAS, & EMOTIONS.
CAN OF WORMS
OPEN AT YOUR OWN RISK!
WARNING! MAY CAUSE EXTREME REACTIONS IN SOME.
WHEN CONSUMED REASONABLY, CAN PROMOTE HEALTH

- **Why should we care?**
  - Algorithms will drive 2 tons of steel on public roads.
  - Some regretful events are statistically bound to happen.
  - Some events will have ethical relevance.

- There are many **ethical viewpoints:**
  - The engineers;
  - The customers;
  - The bystander;
  - The company / companies providing the service;
  - The government;
  - Society at large.

Trolley problems

- The way "trolley problems" are discussed in the popular press has no practical relevance for AV design or policy;
- ...but focusing on an **observable choice** is a great **falsifiable, behavioral approach** to understanding a system.

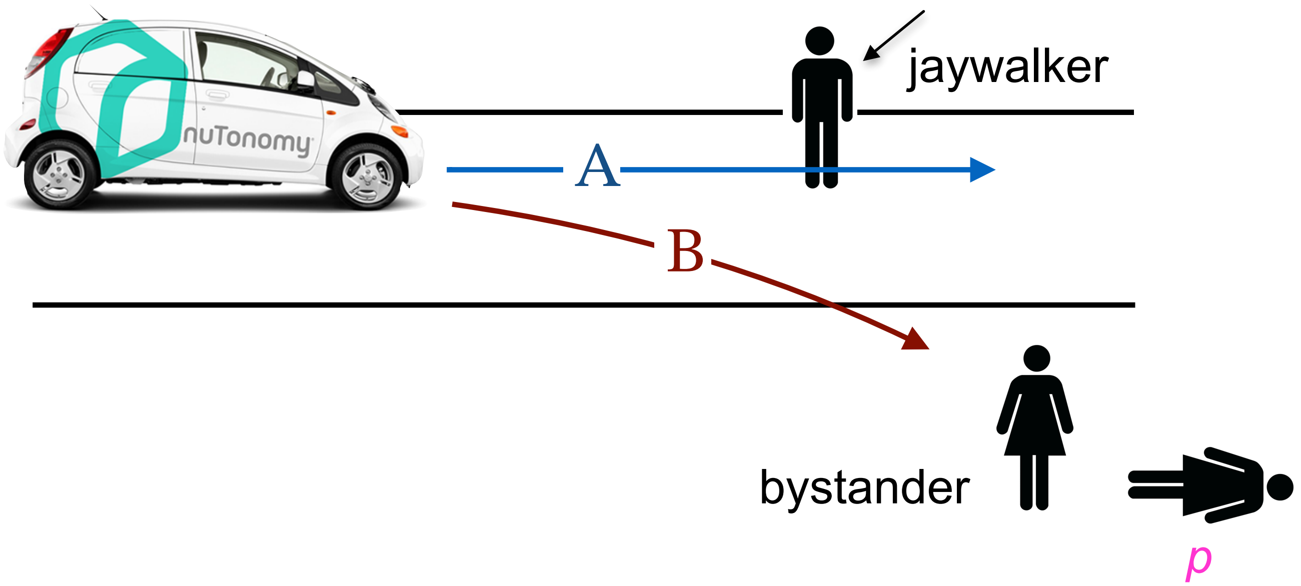


Lin 2016 - Why Ethics Matters for Autonomous Cars
Smith 2016 - The trolley and the Pinto

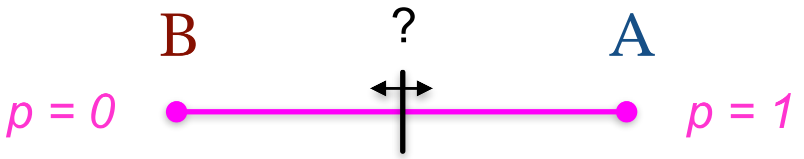
De Freitas 2019 - Doubting driverless dilemmas

The "false positive" trolley problem

- We are driving and a jaywalker walks into the street.
- Suppose our options are:
  - A: We continue on our way, and with probability 1 we hit the jaywalker.
  - B: We swerve, avoiding the jaywalker, but with probability p we hit a bystander.



As a function of p:



Where is the decision threshold?
What does it depend on?

“Right to explanation”

- The **European Union General Data Protection Regulation (GDPR)** (enacted 2016, in effect since 2018), extends the automated decision-making rights in the 1995 Data Protection Directive to provide a legally disputed form of **a right to an explanation.**

  - *“The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.*
    …
    In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.”

Kaminski 2019 The Right to Explanation, Explained

Example: considering bystander age

- Age is an attribute that can be relatively easily estimated and thus could theoretically be used for decision making.

- The **German ministry of transportation** provided official guidelines on ethical decision making for self-driving cars:
  - *"In the event of unavoidable accident situations, **any distinction between individuals based on personal features** (age, gender, physical or mental constitution) is **impermissible**."*

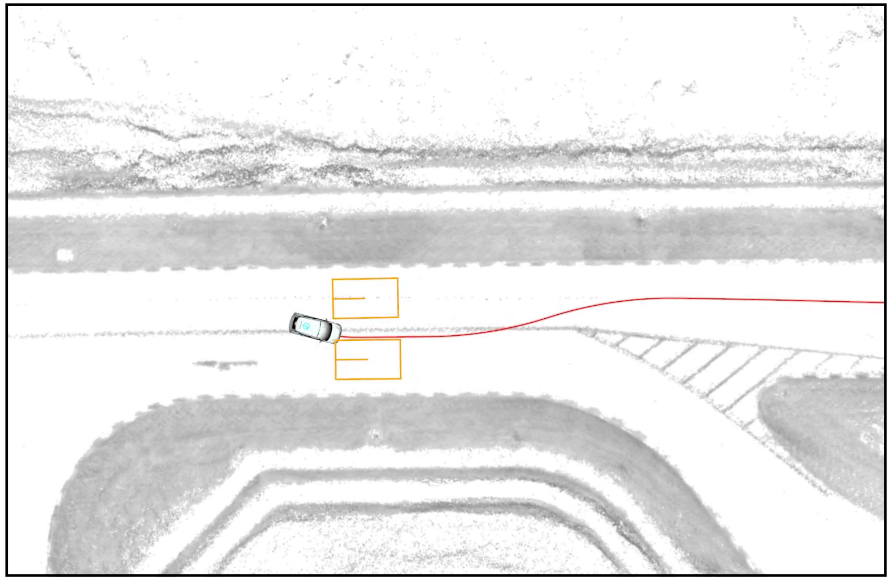
BMVI Ethics Commission 2017 Report

- Value of a Statistical Life:
  - $9.6M / life (U.S. Department of Transportation)
  - $120k per quality-adjusted life-year (QALY) ("dialysis standard")

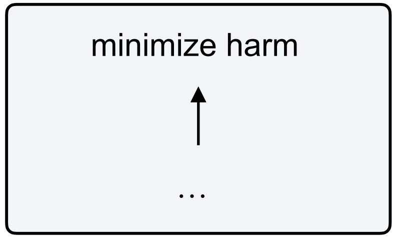
US DOT 2016 report

Example: Liability-aware planning

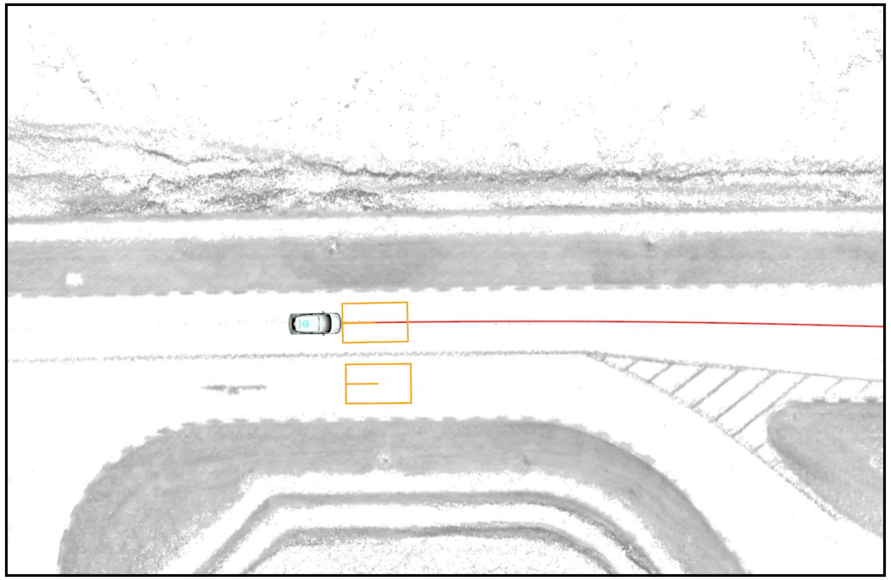
Naive planning:
Minimize harm to humans
(~ minimize kinetic energy transfer)



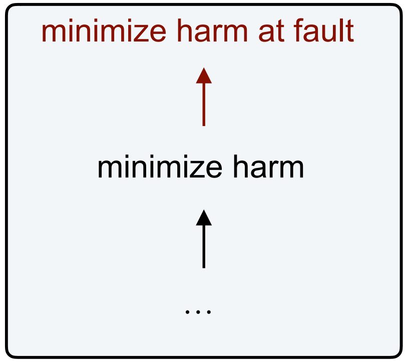
rulebook 1



Liability-aware planning:
Minimize harm to humans
for which we can be blamed.



rulebook 2



What is an engineer to do?

- ❌ Ignore the can of worms
- ❌ Encode your own ethical beliefs
- ✅ Create **transparent** systems
- ✅ Create **customizable** systems
- ✅ **Explain** the issues to the public
- ✅ **Engage with regulators**

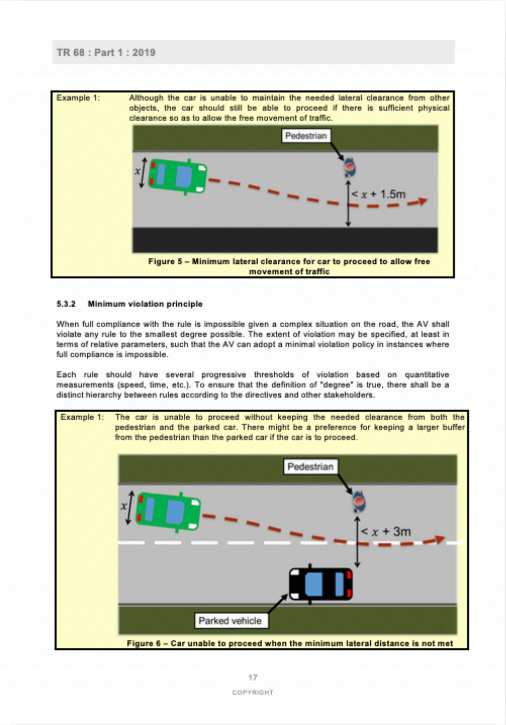
Regulations and informed consent

- **You cannot give meaningful consent if you do not understand what you are consenting to.**
- Measuring AV “safety” is very delicate.
- Regulators just do not have the necessary scientific literacy to understand AVs.

- Pessimistic outlook: in the near future AV safety will continue to be a P.R. topic, fuel for inconsequential speculations of popular press, instead of a serious **discussion from the public health perspective**, which is what would benefit people.

The Singapore example

- Singapore is a virtuous example of regulators being proactive in trying to understand and define regulations in collaboration with academia and industry.
- See” **Singapore’s technical recommendations for AV development**, TR 68, released in January 2019.
- The “minimum violation planning” and “rulebooks” ideas made their way into the TR.



TR 68 : Part 1 : 2019

Example 1: Although the car is unable to maintain the needed lateral clearance from other objects, the car should still be able to proceed if there is sufficient physical clearance so as to allow the free movement of traffic.

Figure 5 – Minimum lateral clearance for car to proceed to allow free movement of traffic

5.3.2 Minimum violation principle

When full compliance with the rule is impossible given a complex situation on the road, the AV shall violate any rule to the smallest degree possible. The extent of violation may be specified, at least in terms of relative parameters, such that the AV can adopt a minimal violation policy in instances where full compliance is impossible.

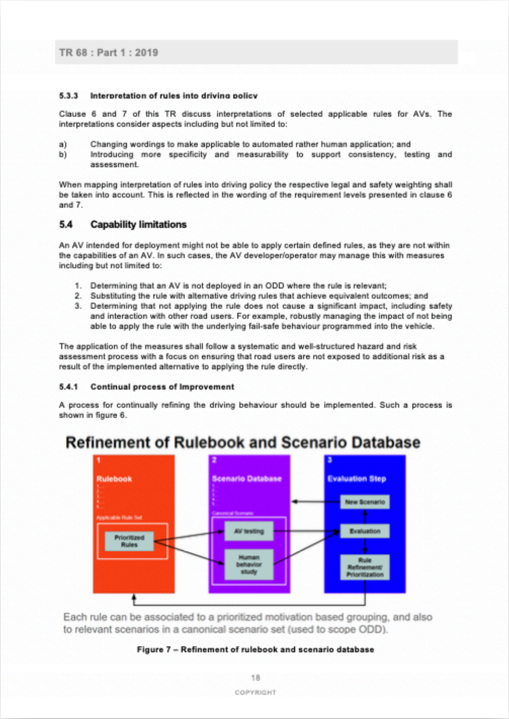
Each rule should have several progressive thresholds of violation based on quantitative measurements (speed, time, etc.). To ensure that the definition of "degree" is true, there shall be a distinct hierarchy between rules according to the directives and other stakeholders.

Example 1: The car is unable to proceed without keeping the needed clearance from both the pedestrian and the parked car. There might be a preference for keeping a larger buffer from the pedestrian than the parked car if the car is to proceed.

Figure 6 – Car unable to proceed when the minimum lateral distance is not met

17

COPYRIGHT



TR 68 : Part 1 : 2019

5.3.3 Interpretation of rules into driving policy

Clause 6 and 7 of this TR discuss interpretations of selected applicable rules for AVs. The interpretations consider aspects including but not limited to:

a) Changing wordings to make applicable to automated rather human application; and
b) Introducing more specificity and measurability to support consistency, testing and assessment.

When mapping interpretation of rules into driving policy the respective legal and safety weighting shall be taken into account. This is reflected in the wording of the requirement levels presented in clause 6 and 7.

5.4 Capability limitations

An AV intended for deployment might not be able to apply certain defined rules, as they are not within the capabilities of an AV. In such cases, the AV developer/operator may manage this with measures including but not limited to:

1. Determining that an AV is not deployed in an ODD where the rule is relevant;
2. Substituting the rule with alternative driving rules that achieve equivalent outcomes; and
3. Determining that not applying the rule does not cause a significant impact, including safety and interaction with other road users. For example, robustly managing the impact of not being able to apply the rule with the underlying fail-safe behaviour programmed into the vehicle.

The application of the measures shall follow a systematic and well-structured hazard and risk assessment process with a focus on ensuring that road users are not exposed to additional risk as a result of the implemented alternative to applying the rule directly.

5.4.1 Continual process of Improvement

A process for continually refining the driving behaviour should be implemented. Such a process is shown in figure 6.

Each rule can be associated to a prioritized motivation based grouping, and also to relevant scenarios in a canonical scenario set (used to scope ODD).

Figure 7 – Refinement of rulebook and scenario database

18

COPYRIGHT

Singapore 2019 - TR 68

The broader picture

- AVs will be remembered as significant because they are the **first tangible application of autonomy.**

- **Welcome to the future:**
  - For the first time **not all “citizens” are humans**.
  - For the first time **we can write laws that are “prescriptive”**.

Conclusions

- **At industrial scale, systems issues dominate** over algorithmic issues.
  - Need: *magic glue*.
  - Need: robust techniques that scale well with complexity.

- **Behavior specification for AVs is “a can of worms”.**
  - Need: transparent, interpretable, customizable systems.
  - Need: better theories for behavior that are compositional and user-friendly.

- **AVs are a proving ground for autonomy.**
  - After AVs, every other application of autonomy looks easy!
  - Public not ready to give “informed consent” due to lack of understanding.