Towards a framework for certification of reliable autonomous systems

Michael Fisher, Viviana Mascardi, Kristin Yvonne Rozier, Bernd-Holger Schlingloff, **Michael Winikoff**, and Neil Yorke-Smith.

Journal of Autonomous Agents and Multi-Agent Systems 35, 8 (2021), 65 pages.

Open access at: https://doi.org/10.1007/s10458-020-09487-2

Towards a framework for certification of reliable autonomous systems

Michael Fisher, Viviana Mascardi, Kristin Yvonne Rozier, Bernd-Holger Schlingloff, **Michael Winikoff**, and Neil Yorke-Smith.

Journal of Autonomous Agents and Multi-Agent Systems 35, 8 (2021), 65 pages.

Open access at: https://doi.org/10.1007/s10458-020-09487-2

Towards a framework for certification of reliable autonomous systems

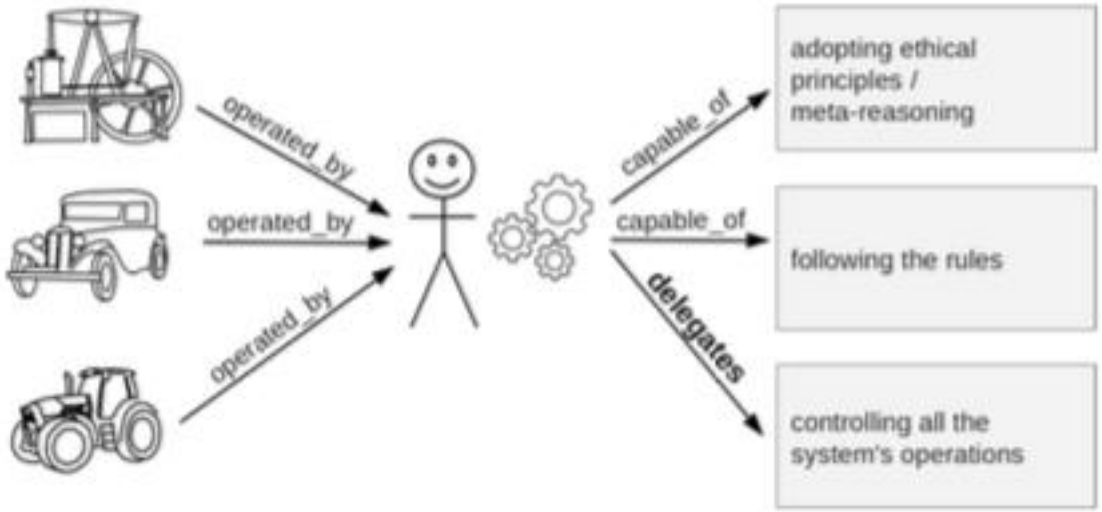
Michael Fisher, Viviana Mascardi, Kristin Yvonne Rozier, Bernd-Holger Schlingloff, **Michael Winikoff**, and Neil Yorke-Smith.

Journal of Autonomous Agents and Multi-Agent Systems 35, 8 (2021), 65 pages.

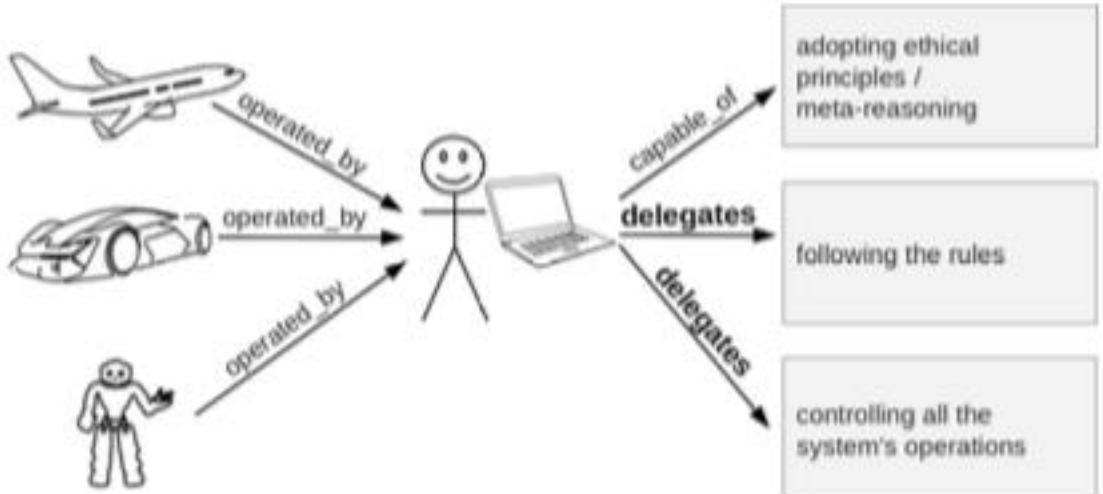
Open access at: https://doi.org/10.1007/s10458-020-09487-2

The challenge: certifying autonomous systems

1:

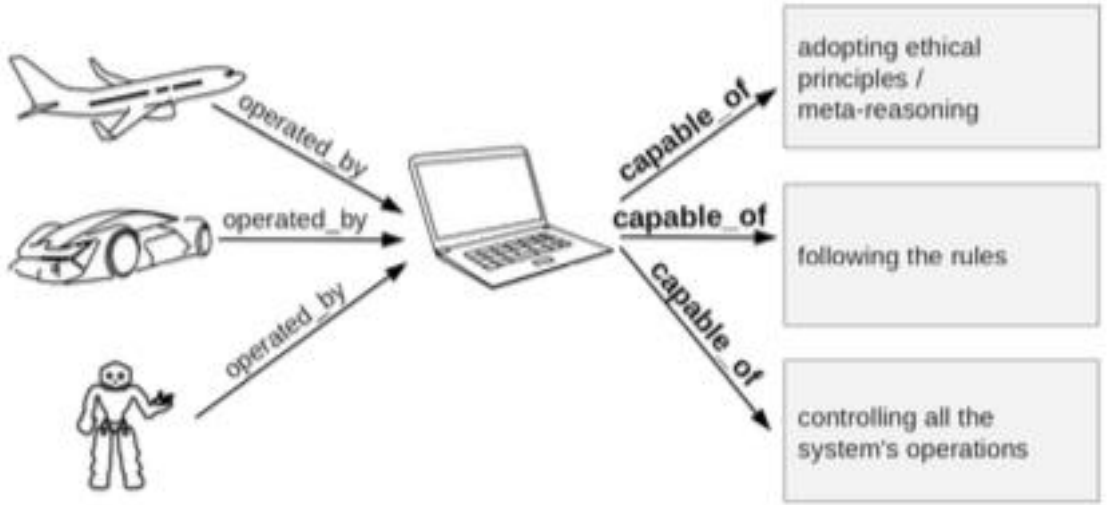


2:



- ... verified ***reliable*** behaviour
- Full autonomy: delegate not just rule following in usual situations, but unusual situations and ethical principles

3:



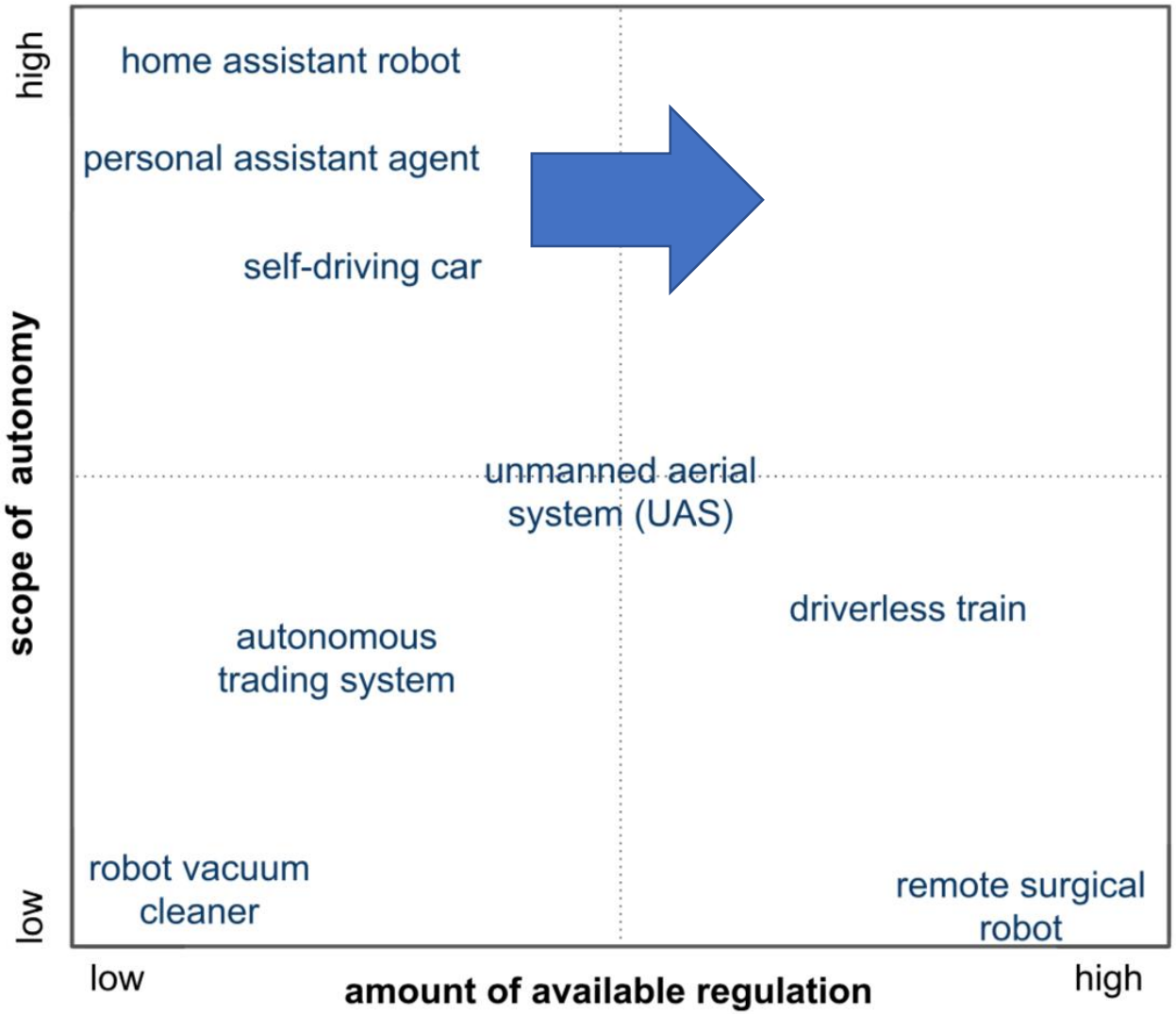
Levels of autonomy (SAE international)

- **No autonomy**
- **Low autonomy**: Straightforward (but non-trivial) tasks are done entirely autonomously (no human poised to take over operation).
- **Assistance systems**: The operator is assisted by automated systems, but either remains in control to some extent or must be ready to take back control at any time.
- **Partial autonomy**: The automated system takes full control of the system, but the operator must remain engaged, monitor the operation, and be prepared to intervene immediately.
- **Conditional autonomy**: The automated system has full control of the operation during specified tasks; the operator can safely turn their attention away but must still be prepared to intervene upon request.
- **High autonomy**: The automated system is capable of performing all planned functions under certain circumstances (e.g., within a certain area); the operator may safely leave the system alone.
- **Full autonomy**: The system can perform all its intended tasks on its own, no human intervention is required at any time.

SAE International (2018). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. https://www.sae.org/standards/conte nt/j3016_201806/

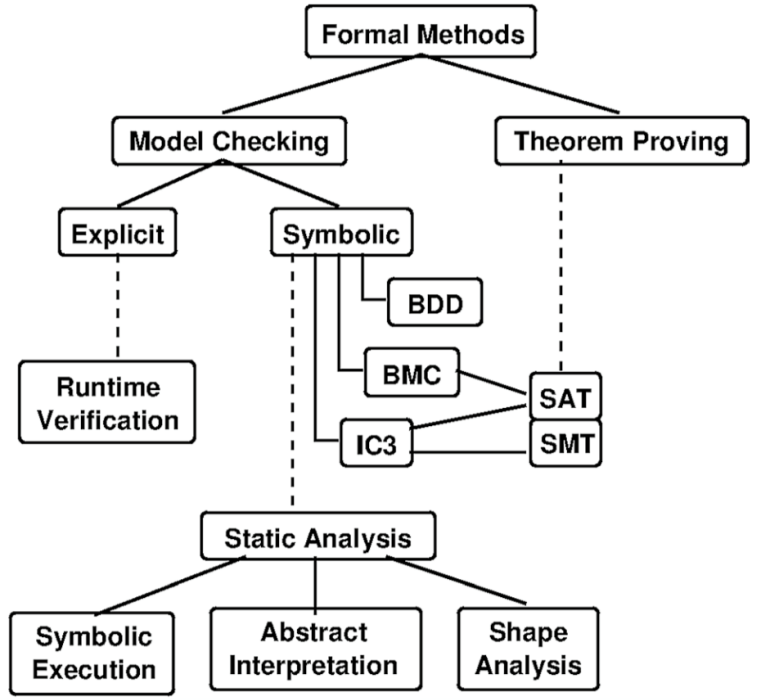
The challenge: certifying autonomous systems

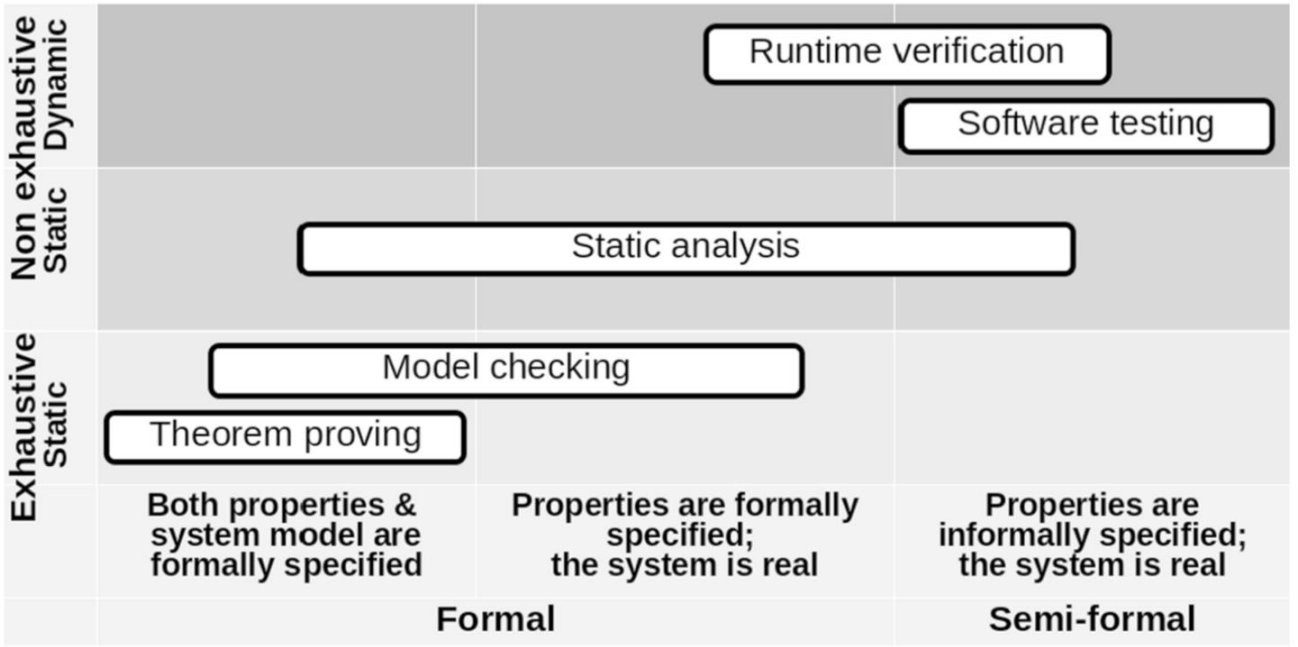
- ... verified ***reliable*** behaviour
- Full autonomy: delegate not just rule following in usual situations, but unusual situations and ethical principles
- Stocktake, framework, and roadmap towards providing verified, reliable behaviour of autonomous systems.



Stocktake

- Standards: many, in different domains, but not (yet) adequate for certifying autonomous systems
- Issues:
  - Requirements: natural/legal text
  - Uncertainty
  - Autonomy
- Techniques:
  - For each method: inputs, outputs, strengths, weaknesses, applicability
  - No silver bullets

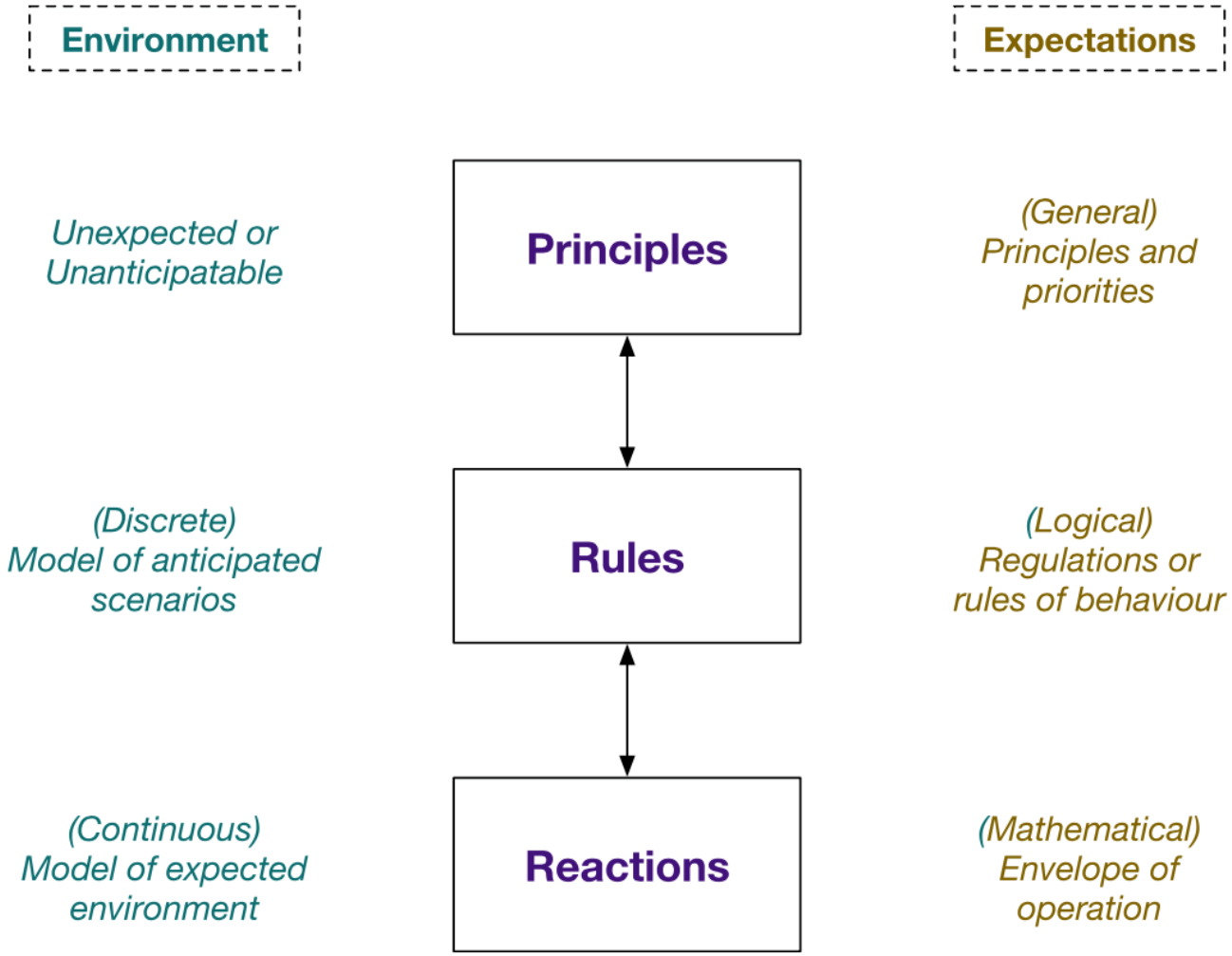




The way forward

1. Verification and validation issues: providing a wide range of techniques, across different levels of formality, that can be used either broadly across the system, or for specific aspects
   - Review of existing techniques and their applicability
2. Architectural/engineering issues: constructing an autonomous system in such a way that it is amenable to inspection, analysis, and regulatory approval
   - A reference three-layer autonomy framework that separates issues
3. Requirements/specification issues: capturing exactly how we want our system to behave, and what we expect it to achieve, overcoming the difficulties arising when human-level rules do not already exist
   - What do we need from regulators, and what process can be used to identify requirements?

Reference three-layer autonomy framework



The way forward

1. Verification and validation issues: providing a wide range of techniques, across different levels of formality, that can be used either broadly across the system, or for specific aspects
   - Review of existing techniques and their applicability
2. Architectural/engineering issues: constructing an autonomous system in such a way that it is amenable to inspection, analysis, and regulatory approval
   - A reference three-layer autonomy framework that separates issues
3. Requirements/specification issues: capturing exactly how we want our system to behave, and what we expect it to achieve, overcoming the difficulties arising when human-level rules do not already exist
   - What do we need from regulators, and what process can be used to identify requirements?

Identifying requirements

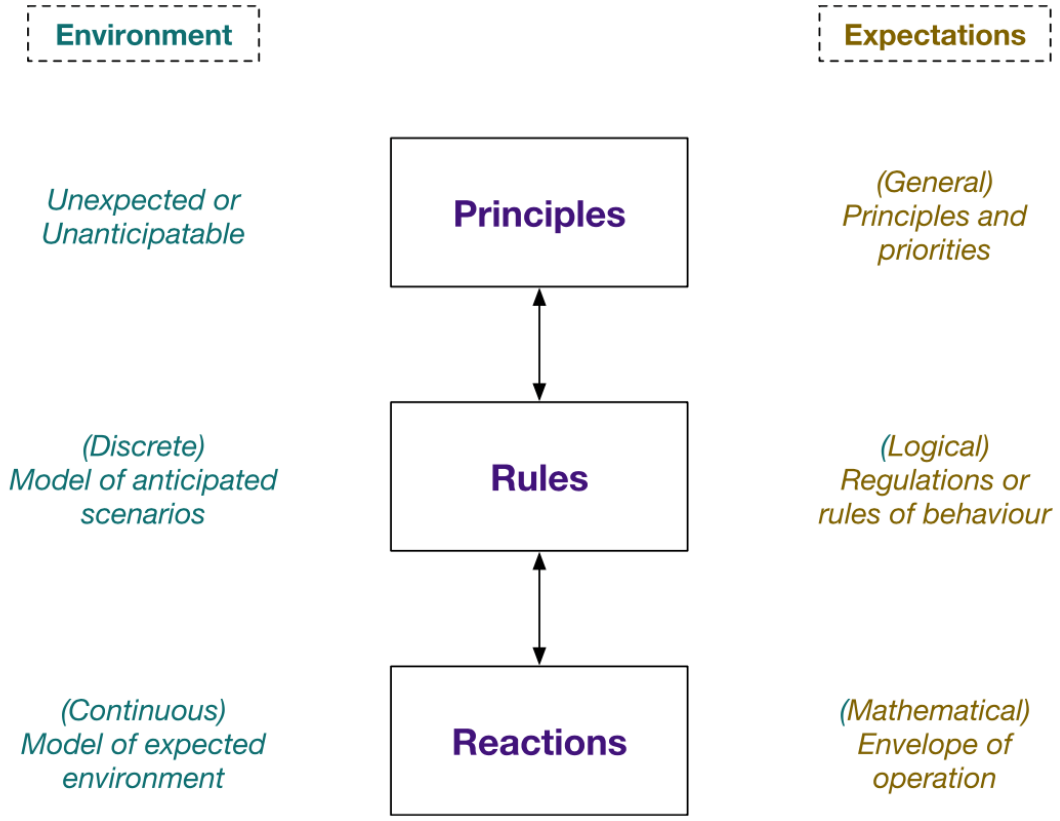
- What is needed from regulators?
  - Not just human-oriented declarative requirements requiring interpretation
  - Do not assume human capabilities and attitudes
  - Include higher-level decision making
  - What does it mean for the system to be safe? What regulations must be followed? What emergent behaviour is anticipated? What would be bad?
  - Assumptions? E.g. system workload
- A process for identifying requirements for certification … (*next slide*)

A process for identifying requirements for certification

- Key idea:
  - Consider the human context (licensing, assumed human capabilities, laws/regulation, the interface) to identify requirements
  - Consider how to assess requirements for an autonomous system; e.g. assess behaviour in scenarios, rather than exam-based assessment of domain knowledge
- Licensing: physical capabilities, domain knowledge, regulatory knowledge, ethical normalisation
- Capabilities: general knowledge, common sense, physical capabilities, maturity/life-experience, ethics/values (e.g. self-preservation)
- Legal/regulation: in what situations should rules be over-ridden?
- Interface: what assumptions does it make about the user? (e.g. ability to reason about a large number of unrelated faults)

Examples to demonstrate our framework

- Positioning wrt layers in the three-layer framework
- Level of (future) autonomy
- Safety criticality
- Available regulation (for autonomous aspects)
- Suitable verification/validation/analysis techniques
- *(example next slide)*



Example: Home assistant robot

- Positioning wrt layers in the three-layer framework: all layers needed, in particular human well-being can pose ethical issues
- Level of (future) autonomy: potentially high; one issue is trust and tradeoff between trust and following rules
- Safety criticality: potentially high (e.g. medication handling, calling for help)
- Available regulation (for autonomous aspects): very little, and falls in the gap (neither industrial robotics nor medical device)
- Suitable verification/validation/analysis techniques: ongoing work, including run-time verification; need to also consider privacy issues

Examples

System	Scope of Autonomy	Targeted future autonomy	Complexity of decision making	Potential harm	Amount of existing regulation
Robot vacuum cleaner	low	high	low	none	low
Autonomous trading system	low-medium	high	low	high	low-medium
Driverless train	low-medium	full	medium	very high	high
Unmanned aerial system	medium	full	high	very high	medium
Self-driving car	medium-high	full	high	high	low-medium
Personal assistant agent	high	full	very high	low-medium	low
Home assistant robot	high	high	very high	medium	low

Future Challenges

Research Challenges

- Specify, perform, and verify principle-based (e.g. ethical) reasoning
- Handling evolving knowledge
- Handling machine learning
- Complexity and scalability
- Synthesising correct systems
- Deriving requirements
- Explanation of systems for regulators

Engineering Challenges

- Assumption provenance
- Building systems that do ethical reasoning
- Managing system synthesis
- Linking verification to the social context: how to generate safety argument?
- High performance verification

Regulatory Challenges

- Multiple stakeholders, including international
- Dealing with industrial secrets
- Compositional verification
- Adequate agreement on what is ‘ethical’?
- Regulating tools
- …

Harel *et al.* (2020)

- Same challenge, focus (like us) on decision making
- Call for research and industry collaboration to develop new foundation
- Highlight three challenges:
    1. Specification of autonomous behaviour, especially given uncertainty (and unanticipated situations)
    2. Dealing with rich environments that can include humans and other autonomous systems – propose to create libraries of environments; also flag identifying behavioural coverage
    3. Building systems by combining ”model-based” and “data-driven” (ML) approaches: requirements, (de)composition, explainability …
- Somewhat more focus on trust and communicating with humans, and more detailed discussion of machine learning
- We propose concrete elements of a solution (three layer framework, process for identifying requirements)

Autonomics: In search of a foundation for next-generation autonomous systems. David Harel, Assaf Marron, Joseph Sifakis
Proceedings of the National Academy of Sciences Jul 2020, 117 (30) 17491-17498; DOI: 10.1073/pnas.2003162117

Contributions

1. proposing a framework for viewing (and indeed building) autonomous systems in terms of **three layers**;
2. showing that this framework is general, by illustrating its application to a range of systems, in a range of domains;
3. indicating what is needed from regulators, and outlining a **process** for identifying requirements; and
4. articulating a range of challenges and future work, including challenges to regulators, to researchers, and to developers.

Acknowledgements

- My co-authors!
- Organisers and participants of Dagstuhl seminar 19112 http://dagstuhl.de/19112
- The anonymous reviewers
- Simone Ancona for the drawings in Section 1
- MF was partially supported by a Royal Academy of Engineering Chair in Emerging Technologies, and by EPSRC grants S4 (EP/N007565), RAIN (EP/R026084), ORCA (EP/R026173), FAIR-SPACE (EP/R026092) and Verifiability Node (EP/V026801).
- KYR was partially supported by NSF CAREER Award CNS-1552934, NASA ECF grant NNX16AR57G and NSF PFI:BIC grant CNS-1257011.
- NYS was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under grant 952215.