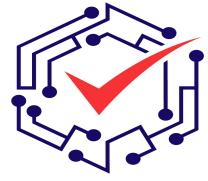
Verifying Autonomous Systems

Michael Fisher

Department of Computer Science, University of Manchester

ORCA/SOLITUDE Workshop *28th September, 2021*

Royal Academy of Engineering



UKRI Verifiability Node

ROBOTICS AND AI IN NUCLEAR

ORCA HUB

Offshore Robotics for Certification of Assets

Remote Safety and Integrity

Assurance and Regulation of Autonomous Robotics

There are relatively few problems, so long as we believe

> the analysis of autonomous systems (robots) involves just a straight-forward application of existing approaches.

Essentially, there's nothing new here!

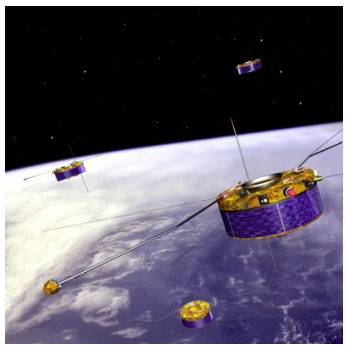
I will highlight some issues with this view:

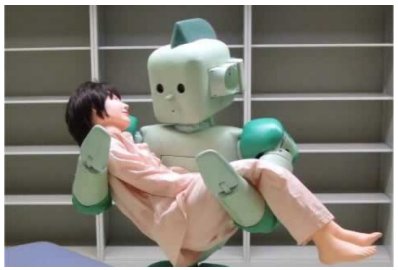
1. autonomy — it *is* different
2. uncertainty — we can't know it all
3. decision-making — not 'what' but 'why'
4. runtime verification — is not (always) the answer

Autonomous Systems

Autonomy:

the ability of a system to make its own decisions and to act on its own, and to do both without direct human intervention.





rtc.nagoya.riken.jp/RI-MAN

www.volvo.com

These often take us *beyond* traditional cyber-physical systems.

Issue: Autonomous Decision-Making

Current approaches to assurance/certification often assume that

- there is a finite set of hazards/failures
- these can be identified beforehand
- this finite set will not change over the life of the system
- .... and so a risk/mitigation based approach can be used

So, in a predictable and known environment
.... we can enumerate all decisions that might be needed and pre-code answers

But: for critical autonomous systems in uncertain environments
.... we need to be clear about how decisions are made
.... crucially, we need to verify **why** *decisions are made*
.... leads us to verifying decision-making process

Who makes the Decisions?

Within 'autonomy' there are important variations concerning decision-making.

Automatic: involves a number of fixed, and prescribed, activities; there may be options, but these are generally fixed in advance.

Adaptive: improves its performance/activity based on feedback from environment — typically developed using tight continuous control and optimisation, e.g. feedback control system.

Autonomous: decisions made based on system's (belief about its) current situation at the time of the decision — environment still taken into account, but internal motivations/beliefs are important.

Distinguishing *between* these variations is often crucial.

Issue: How 'Wrong' are Probabilistic Models?

All probabilistic models of real phenomena/environments are wrong.

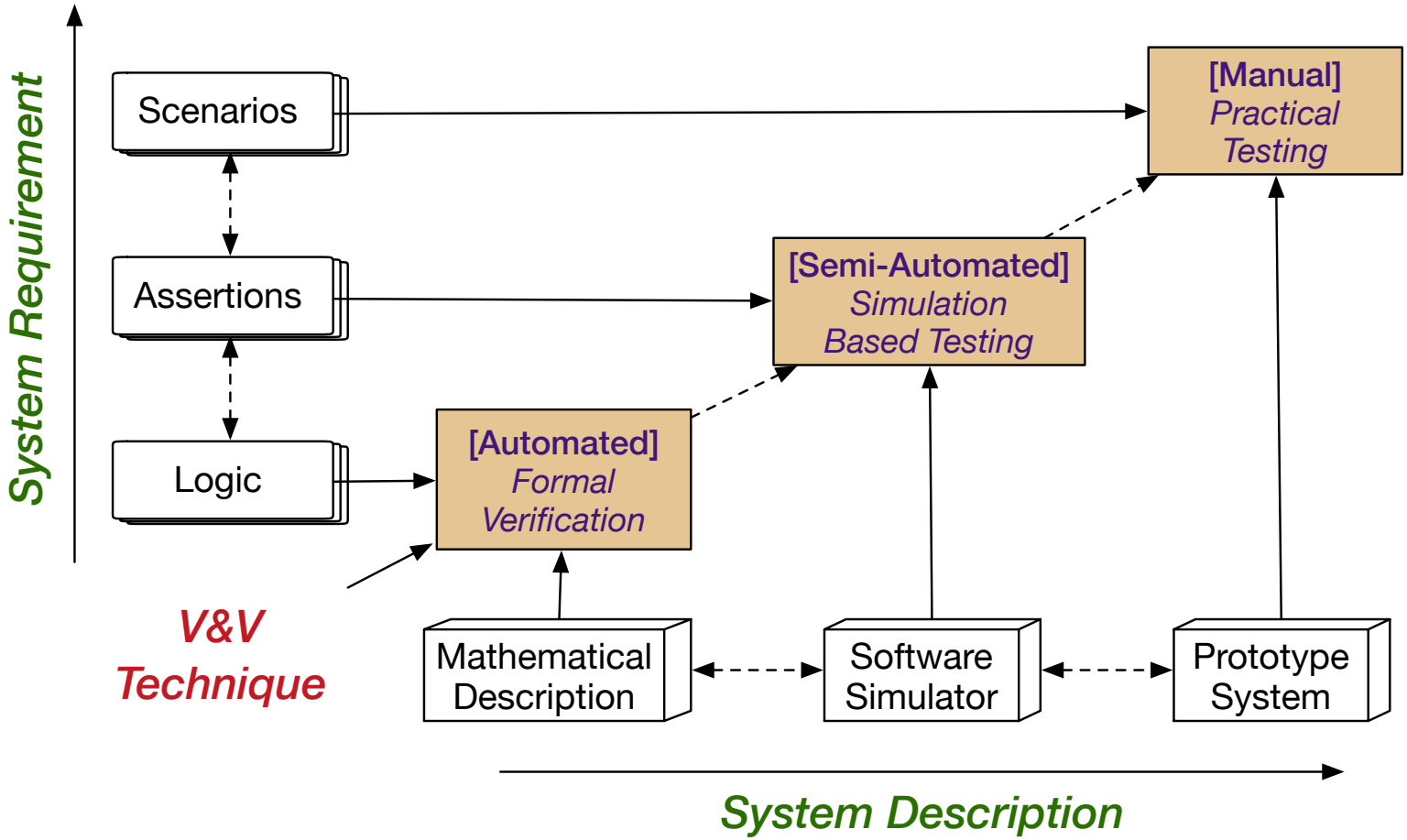
Often very wrong!

Crucially, we don't know how wrong.

When complex feedback control systems (e.g. deep reinforcement learning) are used, even small errors in environmental models can have huge effects.

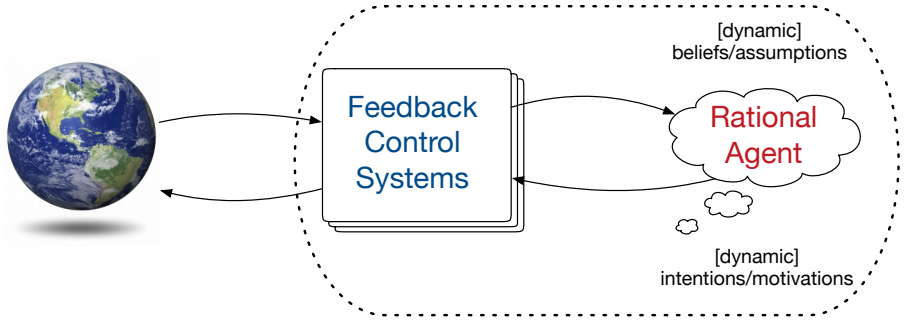
So: evidence we produce using such models is weak.

Use Heterogeneous/Corroborative Verification



[Cite: *A Corroborative Approach to V&V of Human-Robot Teams*]

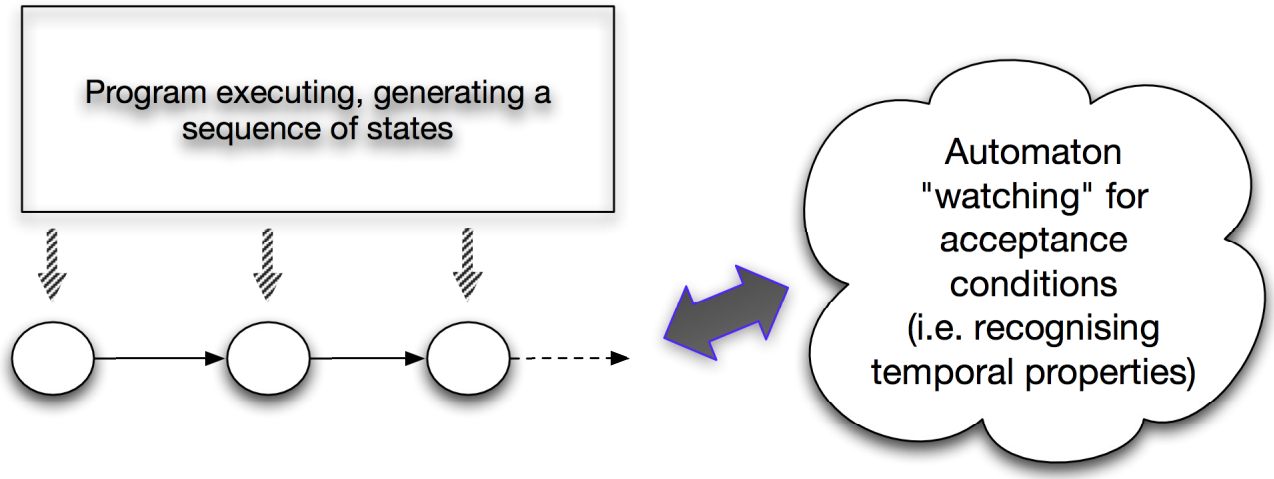
Typical Heterogeneous Verification Approach



Note agent is: symbolic; intentional, transparent, ...

- We might *formally verify* the agent's decision-making
  $\longrightarrow$ we can be certain about this.
- We might *simulate/test* the feedback control components
  $\longrightarrow$ or *verify/monitor*
- We might *practically test* whole system
  $\longrightarrow$ often gives users/regulators more confidence!

Issue: RV isn't (always) the solution



However, the following is far too popular:

1. build an autonomous system, let it do what it wants
2. use RV to flag if it does (or is about to do) something 'bad'

Final Remarks

We should build systems that use components in the right way:

- sub-symbolic AI components, such as machine-learning, for pattern recognition and adaptation;
- symbolic AI components, such as agents, for decision-making and explainability.

Provides Practical hybrid system + Strong evidence for agent decision-making

Important to use heterogeneous/corroborative verificaiton techniques on systems.

Sample Relevant Publications

- Chatila, Dignum, Fisher, Giannotti, Morik, Russell, Yeung. Trustworthy AI. In *Reflections on Artificial Intelligence for Humanity*, 2021.
- Dennis, Fisher, Lincoln, Lisitsa, Veres. Practical Verification of Decision-Making in Agent-Based Autonomous Systems. *Journal of Automated Software Engineering*, 2016.
- Farrell, Luckcuck, Pullum, Fisher, Hessami, Gal, Murahwi, Wallace. Evolution of the IEEE P7009 Standard: Towards Fail-Safe Design of Autonomous Systems. In *International Symposium on Software Reliability Engineering*, 2021.
- Ferrando, Cardoso, Fisher, Ancona, Franceschini, Mascardi. ROSMonitoring: A Runtime Verification Framework for ROS. In *Proc. Towards Autonomous Robotic Systems*, 2020.
- Ferrando, Dennis, Cardoso, Fisher, Ancona, Mascardi. Toward a Holistic Approach to Verification and Validation of Autonomous Cognitive Systems. *ACM Trans. Softw. Eng. Methodology*, 2021.
- Fisher, Cardoso, et al. An Overview of Verification and Validation Challenges for Inspection Robots. *Robotics*, 2021.
- Fisher, Dennis, Webster. Verifying Autonomous Systems. *CACM*, 2013.
- Fisher, Mascardi, Rozier, Schlingloff, Winikoff, Yorke-Smith. Towards a Framework for Certification of Reliable Autonomous Systems. *Autonomous Agents and Multi-Agent Systems*, 2021.
- Luckcuck, Farrell, Dennis, Dixon, Fisher. Formal Specification and Verification of Autonomous Robotic Systems: A Survey. *ACM Computer Surveys*, 2019.
- Webster, Cameron, Fisher, Jump. Generating Certification Evidence for Autonomous Unmanned Aircraft Using Model Checking and Simulation. *Journal of Aerospace Information Systems*, 2014.
- Webster, Western, Araiza-Illan, Dixon, Eder, Fisher, Pipe. A Corroborative Approach to Verification and Validation of Human–Robot Teams. *International Journal of Robotics Research*, 2019.