The Problem with Autonomy

Michael Fisher

Department of Computer Science and
Centre for Autonomous Systems Technology
University of Liverpool

Safety Critical Systems Club, 6th December 2016
Background

Professor of Computer Science

- formal methods, autonomy, proof, programming languages
  
  [http://www.csc.liv.ac.uk/~michael](http://www.csc.liv.ac.uk/~michael)

Director of Centre for Autonomous Systems Technology

- cross-disciplinary centre at University of Liverpool
- involving CS, Engineering, Electronics, Law, Psychology, ...
  
  [http://www.liv.ac.uk/cast](http://www.liv.ac.uk/cast)

Coordinator of EPSRC Network on V&V of Autonomous Systems

- brings together formal verification, testing, user validation, etc
  
  [http://vavas.org](http://vavas.org)

Member of BSI Committees: *Robotic Devices; Robot Ethics*
Why am I here?

- Target *Autonomous Systems* with *significant Autonomy*!

- Show how *Formal Verification* is being used in these
  ... for assessing *safety* (of decisions)
  ... as part of *certification*, and
  ... as a way to potentially analyse *ethics*

- Highlight some of the current problems dealing with such
  *Autonomy*
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.
Even within this, there are 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.
Robotic Assistants are now being designed to help the elderly or incapacitated.

Here is Care-O-bot developed by Fraunhofer IPA.

[ Care-O-bot Promotional Video http://www.youtube.com/watch?v=s9CraxEzZLw ]
Example: Automotive

Road train (‘convoy’) system has control of speed, direction, etc.

Driver can, in principle, take control back.

[ http://www.sartre-project.eu ]
Box Encapsulation demo, NNL Workington:

Large quantities of nuclear material to be sorted, assessed, and disposed of.

Current practice: remote-control of robotic arm.

1. This can be very slow,
2. may involve significant operator expertise (and training), yet
3. is very often mundane, while
4. robot arm cannot be extracted or fixed.
Humans in Charge?

Here is one particular categorisation (PACT):

0: “No Autonomy”  
   → Whole task done by human except for actual operation

1: “Advice only if requested”  
   → Human asks system to suggest options and human selects

2: “Advice”  → System suggests options to human

3: “Advice, and if authorised, action”  
   → System suggests options and proposes one of them

4: “Action unless revoked”  
   4a: System chooses action and performs it if human approves  
   4b: System chooses action and performs it unless human disapproves

5: “Full Autonomy”  
   5a: System chooses action, performs it and informs human  
   5b: System does everything autonomously
Concerns about Full Autonomy

Once the decision-making process is taken away from humans, can we be sure what autonomous systems will do?

Will they be safe? Can we trust them? What if they fail?

Especially important as robotic devices, autonomous vehicles, etc, are increasingly being deployed in safety-critical situations.
Deployed robotic systems typically comprise complex and intricate control systems, neural networks, or genetic algorithms. These can be efficient for optimization tasks, and may be necessary to cope with the continuous nature of real-world interactions, but

- do not capture the essence of autonomy, and
- are often quite hard to analyze and modify
Such hybrid agent architectures are increasingly common across many areas, from software autonomy to autonomous vehicles.
Example: from Pilot to Rational Agent

**Autopilot** can essentially fly an aircraft
- keeping on a particular path,
- keeping flight level/steady under environmental conditions,
- planning route around obstacles, etc.

**Human** pilot makes high-level decisions, such as
- where to go to,
- when to change route,
- what to do in an emergency, etc.

**Rational Agent** now makes the decisions the pilot used to make.
Example: Nuclear Decommissioning Architecture

Key research issues: *vision and sensing; reliable reasoning and decision-making; and robot arm control and grasping.*
Towards Verification

1. *rational agent* for high-level autonomous decisions, and
2. traditional *control systems* for lower-level activities,

These have been shown to be easier to *understand, program, maintain* and, often, much more *flexible*.

We want to formally verify the rational agent within the system — verify the *decisions* the system makes, not its *outcomes*. 
Our Approach: Overall

1. Be clear about *requirements*:
   - autonomy/ethics/trust
   - safety/security
   - functional
   - regulatory/legal requirements

2. Formally verify agent *decisions* → certainty of intention

3. For other components develop high-fidelity simulation and carry out simulation-based testing → HPC

4. Gradually introduce real hardware and real people into test environment in (3)

5. Increasing number of test sites → allows physical testing

6. *...cycle back to (1)...*

7. User validation on (a) simulation, then (b) real system
Robotic Assistants are now being designed to help the elderly or incapacitated.

We can formally verify high-level rules that robot actually uses in deciding what to do.

And so can potentially prove critical properties such as

“robot will always wake the human when it believes there is a fire”
What’s the core difference between a UAV and a manned aircraft?

Obviously: uses a “rational agent” instead of a pilot!

So, why can’t we verify that the “agent” behaves just as a pilot would? i.e. is the agent equivalent to the pilot??

This is clearly impossible, but......
Our Approach

- Autonomous UAS Design/Model
- Formal Logic Specification
- Certification?
- "Model Checking"
- "Abstraction"
- "Selection"
- Rules of the Air

UAS
Rational Agent Controlling the UAV

Our UAV agent has:

- **Beliefs**, for example
  - waiting at runway
  - turning right (e.g. during *sense & avoid*)

- **Desires**, for example
  - complete the mission
  - avoid near-misses

- **Intentions**, for example
  - taxi to runway and hold position
  - turn right to avoid object approaching head-on (i.e. *sense & avoid*), for example
Selected “Rules of the Air”

• “An aircraft shall not taxi on the apron or the manoeuvring area of an aerodrome without [permission]”

• “… when two aircraft are approaching head-on, or approximately so, and there is danger of a collision, each shall alter its course to the right.”

• “[An aircraft in the vicinity of an aerodrome must] make all turns to the left unless [told otherwise]”

Note both the ambiguity and the possible conflict!
UAV has failure → unavoidable crash → but has *some* control

Assesses possible crash sites, but time is running out:

1. on school
2. on field full of animals
3. on a road

System can order options based on *ethical* priorities:

save humans >>> save animals >>> save property
Autonomous system might assess situation and work out that a *serious* accident will occur if human is allowed to take control.

Should system be allowed/expected to ignore human wishes?

Ethical priority likely to be to protect human life over obeying human wishes. But is it legal priority?
Lots of standards; hardly any legislation

- **ISO 10218**: Safety requirements for industrial robots
- **ISO 15066**: Collaborative robots
- ...etc...

Relevant standards from elsewhere: UAV (DO-178C); Automotive (ISO 26262); etc.

None of the above capture “autonomy”!

Relevant standards for autonomy: **BS 8611** Guide to the ethical design and application of robots and robotic systems
Many approaches assume either

- any software in system is essentially simple/deterministic
- there is always a ‘human’ available to take important decisions,
- all hazards/faults/issues can be predicted in advance, or
- the architecture/components of the system remain static.

In many autonomous systems *none* of these are true.
Problems with Autonomy? Humans

- In Social Robotics, *trust* is key
- Disobeying orders will erode trust!
- Privacy violations will erode trust!!
- What if person orders robot to do something robot thinks is *detrimental* to person?
  Or orders robot not to tell anyone?

**Trustworthiness** *versus* human dignity/privacy
*versus* direct human orders
*versus* robot confidence in diagnosis
*versus* legality
*etc*

How do we model/formalise/verify all this?
Problems with Autonomy? Responsibility

**Responsible** — who is accountable, answerable, to blame?

Increasingly, we need *responsible autonomy*:

*We should be certain what the system will decide to do and why it decides to do it*

We are particularly concerned with formal/logical approaches able to *guarantee* this.

**Note:**

*We cannot prove what the system will achieve, since interactions with the real world are always uncertain, but we can prove which decisions the system will make.*

**So:** systems that are *opaque* will not be acceptable!
Key new aspect in Autonomous Systems is that the system is able to *decide for itself* about the best course of action to take.

We have *formal verification* techniques for such autonomous systems.

By verifying the rational agent, we verify not *what* system does, but what it *tries* to do and *why* it decided to try!

However, lots of issues left.....
Thanks to *many* people.....

The work described in this talk is due to *many* people.....

- Louise Dennis (Computer Science, Univ. Liverpool)
- Matt Webster (Computer Science, Univ. Liverpool)
- Clare Dixon (Computer Science, Univ. Liverpool)
- Rafael Bordini (UFRGS, Brazil)
- Maryam Kamali (Computer Science, Univ. Liverpool)
- Alexei Lisitsa (Computer Science, Univ. Liverpool)
- Sandor Veres (Engineering, Univ. Sheffield)
- Owen McAree (Engineering, Univ. Sheffield)
- Mike Jump (Engineering, Univ. Liverpool)
- Richard Stocker (NASA Ames Research Center, USA)
- Marija Slavkovik (Univ. Bergen, Norway)
- Alan Winfield (Bristol Robotics Lab)

- EPSRC, for funding many of these activities.
Sample Relevant Publications


