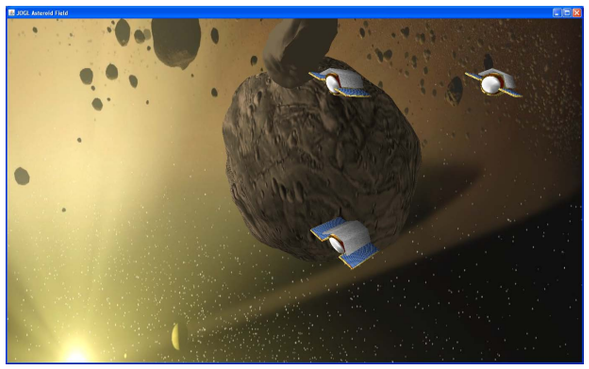
Michael Fisher

Trusting Autonomy: Verification and Certification

What is the Problem?

Core aspect of **AUTONOMY**:

autonomous systems make decisions (and take actions) ***without*** *human intervention*

But:

- can we be sure what decisions they will make?
- can we be sure what actions they will then take?
- (most importantly) can we be sure ***why*** they make these choices?

Trustworthiness

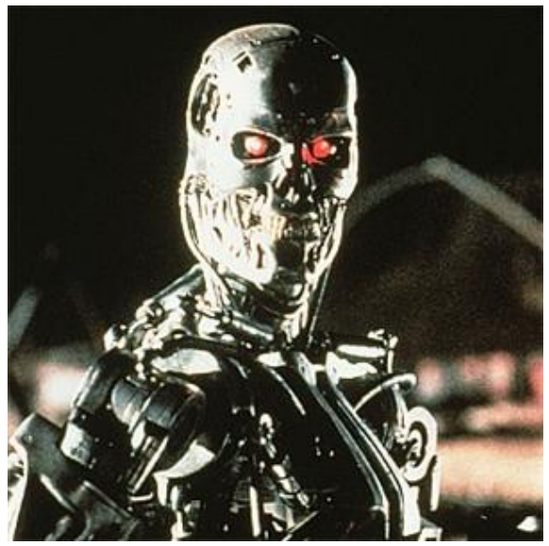
There are (at least) two key elements to trusting autonomous systems:

- do we trust that the system is *reliable*?

  will it always work correctly?

- do we trust that the system has the right *intention*?

  is it always working for our benefit?



If we do not know when, how, and **why** these systems make their decisions then we will not trust them.

We need ***strong*** verification to help convince us - testing/sampling is not enough

Without this ⟶ **Regulators** will/should not certify them or allow them to be used

Excuses?

Strong verification (especially formal verification)

- requires too much effort - analysts need much more expertise than is common
- is too complex - formal verification takes much too long
- is not possible - engineered system is too complex/opaque

No Psychiatrists for Robots?

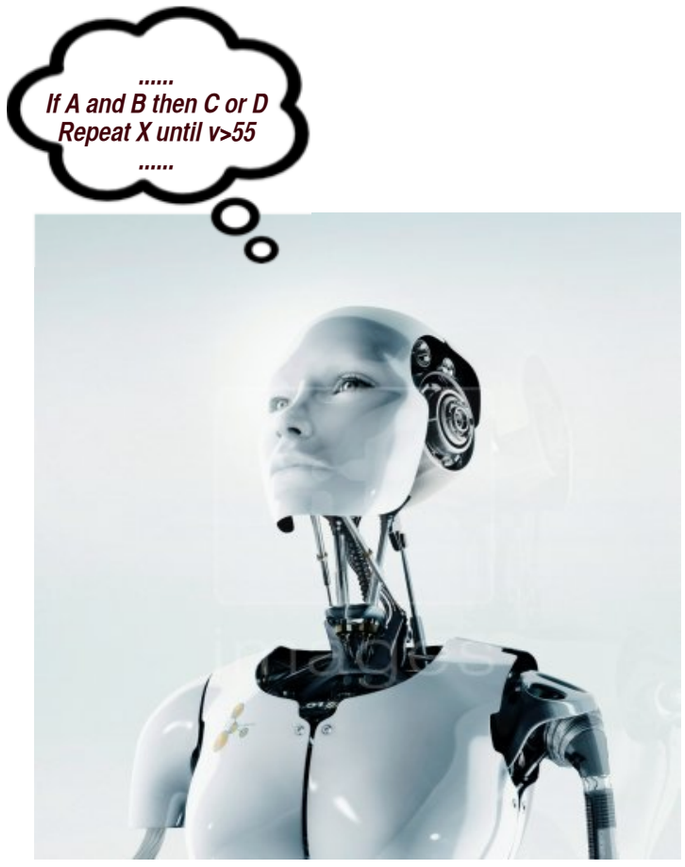
Since we built the autonomous system we can (in principle)

examine a system's internal programming and expose exactly

1. *what* it is `thinking'
2. what *choices* it has, and
3. why it *decides* to take particular ones.

To ensure this we capture high-level decision-making in symbolic components.

In this way we can expose the reasons for decisions and can ***formally verify*** that its decisions will always be taken for the right reasons.



Verification ⟶ confidence, trust, certification

Once we can ***expose*** **why** a system makes its decisions then this:

⟶ can help convince the **Public** that the system has “good intentions”

- we can **explain** (and **record** - “ethical black box”) what it does, and why

- we can match these intentions against societal **ethics/norms**

⟶ can help convince **Regulators** to allow/certify these systems

- we can **verify** (prove) that it always makes decisions in an appropriate way

- we can **verify** (test) reliability of learning/adaptation/actuation components

⟶ can give **Engineers** confidence to build truly autonomous systems

Message:

Architect your systems **well** to expose intentions/reasons and provide strong **verification** for crucial decision-making components ⟶ *trustworthy autonomy*

Example: **Trustworthy Autonomous Systems**

See:
epsrc.ukri.org/newsevents/events/ukri-trustworthy-autonomous-systems-programme-town-hall-meeting
epsrc.ukri.org/funding/calls/trustworthy-autonomous-systems-research-nodes-call
epsrc.ukri.org/files/funding/calls/2019/trustworthy-autonomous-systems-hub-full-proposal