AAAI Spring Symposium; 23rd March 2022

PANEL: Assured trust and autonomy in human machine teams

Michael Fisher
[Department of Computer Science, University of Manchester, UK]

https://web.cs.manchester.ac.uk/~michael

Royal Academy of Engineering

MANCHESTER 1824
The University of Manchester

UKRI Verifiability Node

Teams and Trustworthiness

Teams are social constructs
→ so using high-level agent approaches (for modelling/programming) makes sense
→ agents; intentions; beliefs; goals; responsibilities; context; collaboration; etc.

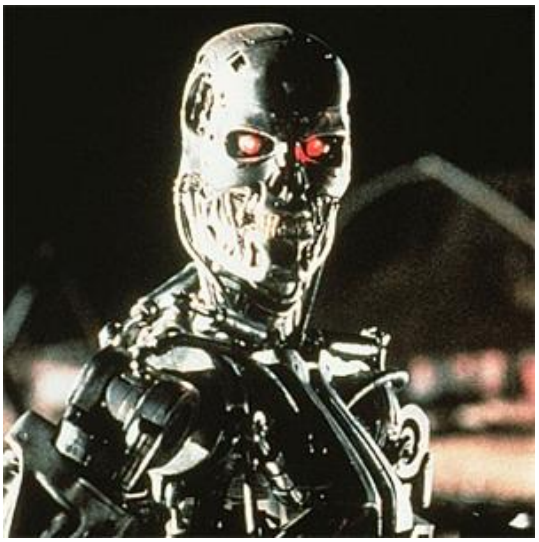
Teams work better when members trust each other

Trust is subjective, but systems can be designed to be **trustworthy**

Trustworthiness in standard Cyber-Physical Systems essentially equates to *reliability*

Trustworthiness in Autonomous Systems involves *reliability* plus *beneficiality*

What would make this Trustworthy?



Obviously, looks don't help here, but

BENEFICIALITY:
- Transparency - especially of intention
- Verifiability - so that we can prove it will never try to do anything `bad'

RELIABILITY is here much less of an issue

Trust - what can we do before deployment?

Systems can be designed to be **trustworthy**

Engineering reliability is important, but it is vital we have:

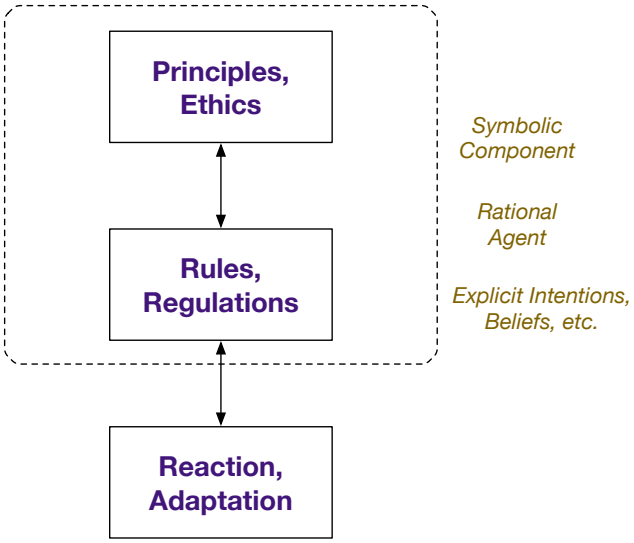
1. Transparency of behaviour/intention, and so explainability and (ideally) understanding

3. Strong Verifiability, specifically formal verification, and certainty about decisions/ intentions

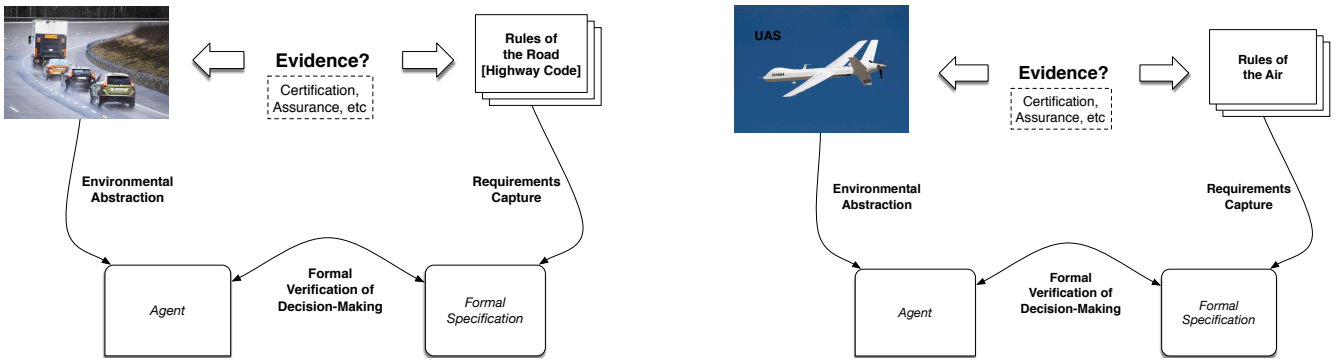
For individual robots/vehicles we strongly verify key components (e.g. decision-making) - we never *rely* on anything that has not been formally verified!

Similarly with teams - would you trust an opaque teammate?

Clarity over Decision-Making



Verifying Conformance to Human-level Rules



Verifying Ethical Reasoning

Once the agent decisions take ***ethical concerns*** into account then we can extend our formal verification to also assess these.

Capturing ethical requirements is difficult - cf Philosophy

These can range from

SIMPLE ORDERING OF CONCERNS

save life >> save animals >> save property

to

(FULL) ETHICAL THEORIES

Complexity issues!

Some Examples

Advice: providing legal/ethical advice - e.g. firefighting

- Explain laws/ethics, based on situation
- Guarantees of correctness concerning advice given

Social/Domestic Robotics: e.g. care robots, social robots

- Reliability/predictability, especially physical interaction
- Exposure of intention → gain trust
- Verification of truthfulness → maintain trust
- Match human ethics, e.g: “don’t tell anyone when I do this”

Space Robotics, e.g. planetary habitat checking and remote/ infrequent human interaction

- Need autonomy, as issues can’t all be predicted
- Verification of the decision-making processes
- Need resilience and reconfigurability
- Clearly explain pertinent issues to operators/astronauts

Building Teams?

Reliability is important in a team member

Beneficiality is even more important once autonomous entities are involved
- transparency → leading to explainability and verifiability

We must
1. be sure what the motives/intentions of the autonomous team members are
2. be able to prove that a team member will never (deliberately) do something `bad'