Dependability issues in AI-based systems

David Powell, Pascale Thévenod-Fosse
LAAS-CNRS
Context: autonomous space systems

- Autonomous systems
  - Classic automated systems
  - AI-based systems

- Literature search
  - Standards
  - Critical AI-based systems
Main sources


Knowledge representation

- Advantages
  - re-usability
  - maintainability
  - understandability

- But
  - style adapted to inference mechanism
  - tighter links than one would expect from theory
  - intertwining of declarative and procedural knowledge can give rise to inconsistencies
Inference System function

- Logical consistency (determined by designer)
  - Deductive theorem prover (guaranteed)
  - Production rule system (not guaranteed)
- Logical consistency (emergent from examples)
  - Inductive learning (guaranteed)
  - Neural network (not guaranteed)

[Robertson & Fox 2000]
Inference

System function

determined by designer

determined by designer

Logical consistency

guaranteed

guaranteed

Deductive theorem prover

Production rule system

Logical consistency

not guaranteed

not guaranteed

Inductive learning

Neural network

Examples

[Robertson & Fox 2000]

safety argument: predicted behavior
Inference

System function

determined by designer

Logical consistency

guaranteed

Deductive theorem prover

Safety argument: predicted behavior

Examples

not guaranteed

Production rule system

Formal proof

Emergent from examples

Logical consistency

guaranteed

Inductive learning

not guaranteed

Neural network

[Robertson & Fox 2000]
Inference

System function

Logical consistency
- determined by designer
- guaranteed
- Deductive theorem prover

Production rule system
- not guaranteed

Logical consistency
- emergent from examples
- guaranteed
- Inductive learning

Logical consistency
- not guaranteed
- Neural network

Examples

- [Robertson & Fox 2000]

Safety argument: predicted behavior
- formal proof
- code structure
Inference

System function

- Logical consistency determined by designer
- Logical consistency emergent from examples

Examples

- Deductive theorem prover: guaranteed
- Production rule system: not guaranteed
- Inductive learning: guaranteed
- Neural network: not guaranteed

[Robertson & Fox 2000]

Safety argument: predicted behavior
Formal proof
Code structure
Formal proof
Inference

System function

determined by designer

Logical consistency

guaranteed
Deductive theorem prover

not guaranteed
Production rule system

Logical consistency

guaranteed
Inductive learning

emergent from examples

not guaranteed
Neural network

[Robertson & Fox 2000]
safety argument: predicted behavior
formal proof
code structure
formal proof
testing
Verification of rule bases

- **Redundancy**
  - same premises and same effects
- **Conflicts**
  - same premises and different effects
- **Subsumptions**
  - same effects but one premise more restrictive
- **Unnecessary conditions**
  - same effects with contradictory conditions in premises, e.g., X and not(X)
Deep Space One  [Muscettola et al. 1998]

• Autonomous spacecraft
  - Very restricted human intervention: decision-making with strict deadlines & resource constraints; HW-fault tolerance; concurrent activities (conflicts)
  - First AI-based spacecraft: model-based programming, on-board deduction and goal-based closed loop control

• Lessons
  - schedule impacts: early need to encode knowledge
  - human-centered operations: autonomy has limits
  - validation & testing major concern:
    • ability to focus on domain model does help
    • cost-effective extensive testing requires an automated test oracle
Agent technology in medicine [Fox & Das 2000]

Knowledge-based medical decision support

- Problem goals
- Situation beliefs
- Actions
- Candidate solutions
- Decisions
- Plans

1. Problem definition
2. Propose solutions
3. Argue
4. Commit
5. Commit
6. Schedule
7. Data acquisition

Agent technology in medicine [Fox & Das 2000]

Knowledge-based medical decision support

- **Problem goals**
- **Situation beliefs**
- **Actions**
- **Candidate solutions**
- **Decisions**
- **Plans**

- Extension to automatic decision commitment:
  - current knowledge cannot lead to different best action
  - cost of more knowledge > cost of inappropriately committing current preference
Agent technology in medicine [Fox & Das 2000]
Agent technology in medicine [Fox & Das 2000]

- Formal methods (specification, description, proof...) and rigorous testing good, but...
Agent technology in medicine [Fox & Das 2000]

- Formal methods (specification, description, proof...) and rigorous testing good, but...
- Insufficient for critical systems that must operate in an open environment that:
  - cannot be fully monitored or controlled
  - in which unpredictable events will occur
Agent technology in medicine [Fox & Das 2000]

- Formal methods (specification, description, proof...) and rigorous testing good, but...
- Insufficient for critical systems that must operate in an open environment that:
  - cannot be fully monitored or controlled
  - in which unpredictable events will occur

  safety bag
Agent technology in medicine [Fox & Das 2000]

• Formal methods (specification, description, proof...) and rigorous testing good, but...

• Insufficient for critical systems that must operate in an open environment that:
  - cannot be fully monitored or controlled
  - in which unpredictable events will occur

  safety bag

• Domain-independence?:
  - guardian agent concept for active safety management
  - safety logic: includes deontic modalities such as authorized, preferred, permitted and obligatory
Hazards of AI-based systems
Hazards of AI-based systems

- Technical factors
Hazards of AI-based systems

- Technical factors
  
  $H_1$: knowledge "wrong": incorrect belief, missing data
Hazards of AI-based systems

• Technical factors
  
  H1: knowledge “wrong”: incorrect belief, missing data
  H2: unsound inference: procedures incorrect in some way
Hazards of AI-based systems

- Technical factors
  - H1: knowledge “wrong”: incorrect belief, missing data
  - H2: unsound inference: procedures incorrect in some way
  - H3: unforeseen contingencies: unusual circumstances
Hazards of AI-based systems

- **Technical factors**
  - H1: knowledge “wrong”: incorrect belief, missing data
  - H2: unsound inference: procedures incorrect in some way
  - H3: unforeseen contingencies: unusual circumstances
  - H4: specificity of decision criteria: not universally acceptable
Hazards of AI-based systems

• Technical factors
  H1: knowledge “wrong”: incorrect belief, missing data
  H2: unsound inference: procedures incorrect in some way
  H3: unforeseen contingencies: unusual circumstances
  H4: specificity of decision criteria: not universally acceptable

• Human factors
Hazards of AI-based systems

• Technical factors
  H1: knowledge “wrong”: incorrect belief, missing data
  H2: unsound inference: procedures incorrect in some way
  H3: unforeseen contingencies: unusual circumstances
  H4: specificity of decision criteria: not universally acceptable

• Human factors
  H5: ontological mismatch: meaning of term — system vs user
Hazards of AI-based systems

- **Technical factors**
  - H1: knowledge “wrong”: incorrect belief, missing data
  - H2: unsound inference: procedures incorrect in some way
  - H3: unforeseen contingencies: unusual circumstances
  - H4: specificity of decision criteria: not universally acceptable

- **Human factors**
  - H5: ontological mismatch: meaning of term — system vs user
  - H6: overconfidence of user wrt knowledge-based system
Hazards of AI-based systems

• Technical factors
  H1: knowledge “wrong”: incorrect belief, missing data
  H2: unsound inference: procedures incorrect in some way
  H3: unforeseen contingencies: unusual circumstances
  H4: specificity of decision criteria: not universally acceptable

• Human factors
  H5: ontological mismatch: meaning of term — system vs user
  H6: overconfidence of user wrt knowledge-based system
  H7: incredulousness: e.g., due to no explanation of system reasoning
Conclusions 1/2
Conclusions 1/2

• Key is separate knowledge representation (H1, H2)
  - more readily checkable by domain experts
  - consistency & completeness of knowledge base
  - V&V of knowledge-independent components similar to classical software engineering
Conclusions 1/2

• Key is separate knowledge representation (H1, H2)
  - more readily checkable by domain experts
  - consistency & completeness of knowledge base
  - V&V of knowledge-independent components similar to classical software engineering

• Learning systems (H1, H2, H3)
  - prove to be quite robust in practice
  - but less amenable to dependability arguments
Conclusions 1/2

• Key is separate knowledge representation (H1, H2)
  - more readily checkable by domain experts
  - consistency & completeness of knowledge base
  - V&V of knowledge-independent components similar to classical software engineering

• Learning systems (H1, H2, H3)
  - prove to be quite robust in practice
  - but less amenable to dependability arguments

• Early encoding of domain-specific knowledge (H1)
  - progressive refinement using evolutionary program development strategy
Conclusions 2/2
Conclusions 2/2

• Challenge of autonomous decisional systems (H3, H4)
  - should act sensibly in unanticipated and complex situations
    • confidence building through extensive testing
      - envisage automated test oracle
    • on-line assurance (fault tolerance)
      - safety bag to avoid catastrophic failures
      - generalization towards active safety management?
Conclusions 2/2

• Challenge of autonomous decisional systems (H3, H4)
  - should act sensibly in unanticipated and complex situations
    • confidence building through extensive testing
      - envisage automated test oracle
    • on-line assurance (fault tolerance)
      - safety bag to avoid catastrophic failures
      - generalization towards active safety management?

• Human intervention (H5-H7)
  - need to recognize that it may be necessary
  - interaction between humans and AI-based systems introduces new human factor risks: ontological mismatch; overconfidence; incredulousness...