Risk, Uncertainty and AI

Non-probabilistic methods for anticipating and preventing AI risks

Prof. Vicki Bier - U of Wisconsin
Prof. A Sasha Gutfraind - Loyola University Chicago
October 2023

Outline

1. Motivation
2. Risk-reducing Design and Operations Toolkit (RDOT) for risk management
3. Applying RDOT to AI safety engineering

Motivation

1. AI is being implemented in more applications every day, often by non-experts

2. It is challenging to apply decision theory to manage AI risk due to:
    a. Unknown space of possible events
    b. Difficulty in estimating probabilities
    c. High complexity of analysis

What if there was an inexpensive but proven way to manage AI risks…?

Background: Classic approaches to uncertainty

1. Decision theory: select among alternatives based on expected utility (Savage)
2. Decision heuristics: greedy, minimalist & others (Gigerenzer & others)
3. HERE: risk-reducing design and operations toolkit (RDOT)

RDOT: strategies for decision under uncertainty and uncertainty reduction

- Prompt: “how are risks or uncertainty (of different varieties) managed here”
- Curated from engineering, business, medical, and others literatures
- 90+ strategies were identified (Gutfraind 2023) https://doi.org/10.5281/zenodo.8350550

Types of RDOT strategies

1. **Configurational strategies that design or improve preparedness for uncertainty:**
   a. Robustness (e.g., factor of safety)
   b. Defense in depth
   c. Compartmentalization
2. **Reactive strategies that improve detection of events and subsequent responses:**
   a. Anomaly detection
   b. Standoff interdiction
   c. Incident response units
3. **Formal strategies involving algorithms or workflows:**
   a. System simulation
   b. Hypothetico-deductive method
   c. Hazards and operability studies (HAZOPS)
4. **Cross-cutting strategies for special situations:**
   a. Adversarial strategies, beneficial uncertainty, strategies that enable future flexibility

Case study: Chatbot Q&A solution for a retailer

A small software developer uses a knowledge base + language model to build a customer service chatbot

Known risks: hallucinations; off-topic discussion; privacy of training data; malicious jailbreaking

Existing solution strategies:

- Filtering of inputs and outputs
- Prototype-driven development
- Pre-release testing
- Post-release monitoring
- Certain domain-specific measures (Wei et al., 2023, arxiv.org/abs/2307.0248)

RDOT strategies: configuration

- Fail-safe design
- Multi-layer defense
- “Stop button”
- *Safety culture*

RDOT strategies: detection and reaction

- Anomaly detection system
- Incident response unit

RDOT strategies: formal methods

- HAZOP
- Independent certification
- Incident investigation

Case study: autonomous vehicles

The year is 2030, and a software developer applies a newly-introduced “vehicle AI” to move and operate agricultural equipment

Known risks: loss of control, violation of safety, collaboration with humans

Existing solution strategies

- Redundant multi-spectrum sensors
- Limit to small / low impact vehicles
- Testing in simulation/controlled conditions
- Standby human driver
- 3rd party software audit

RDOT solutions
(letters A through E)

Accelerate adaptation
Adjust planning horizon
Anomaly detection and investigation
Automatic containment system
Basic research
Blue/green deployment
Canary detection
Contingency planning
Coordinate action
Decision template
Deflect
Delay
Delegation and local empowerment
Dispersed storage
Early warning system
Eliminate input variables
Escapable design
Exhaustive analysis of all actions
Event forensics and attribution
Event tree analysis
Evolutionary architecture
Expansive analysis
Expert elicitation and judgment

Allocating resources between strategies

- In classical decision theory we select *i*

  $$\max_i \; E[u_i]$$

- Radical uncertainty makes calculating expectation difficult / intractable

- We therefore could:
  1. Select strategies based on familiarity, convenience, standards/regulations
  2. Risk and control matrix (RACM) framework
  3. Using proxy metrics
  4. Multi-objective proxy measures

Multi-objective proxy measures for RDOT

If the outcomes of interest are

$$\max_x \{K_1(x),\ ..,K_p(x)\}\ \ such\ that\ P(x)\ \le\ Q$$

We add proxy metrics (e.g. resilience)

$$\max_x \{K_1(x),\ ..,K_p(x), L_1(x),..,L_q(x)\}\ \ such\ that\ P(x)\ \le\ Q$$

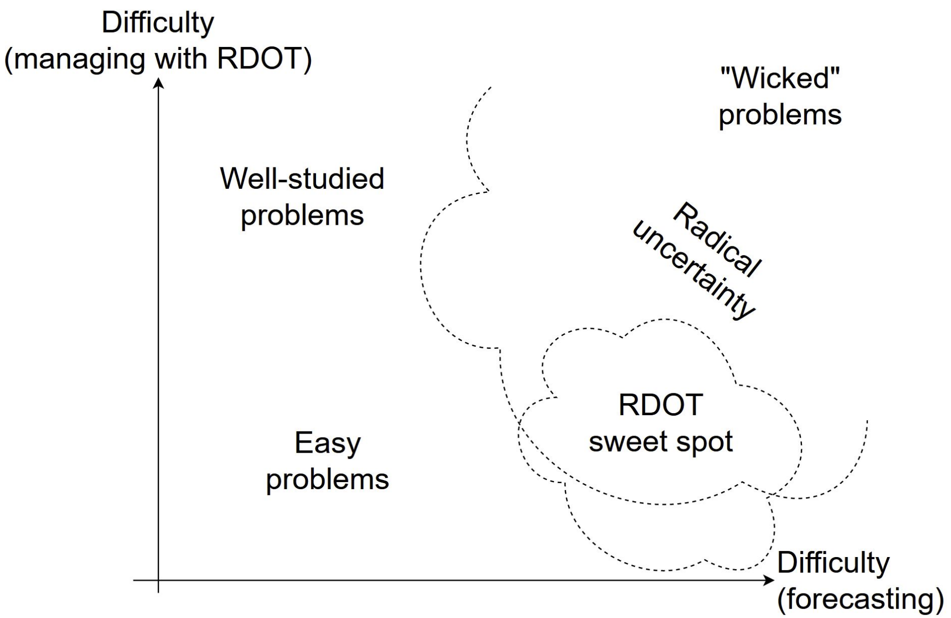
Addressing radical uncertainty with RDOT

- Important problems affected by uncertainty can be manageable:

  e.g., Columbus crossed an ocean by using well-built ships and trained crews lacking any understanding of fluid dynamics or geophysics

  e.g., multi-layer defense works for AI even when we can't forecast the risks

- Some problems are easy to RDOT but hard for decision theory
- "Wicked" might remain hard



arxiv.org/abs/2301.10244v1

Strengths and weaknesses of RDOT

Strengths

- Natural solution for many problems
- Readily utilizable by non-experts
- Does not require complex estimation
- Could address emergent or poorly-understood risks
- Transferable across application areas

Weaknesses

- Satisficing rather than optimizing solutions
- Multiple overlapping strategies create new modes of failure and inefficient designs
- Difficult to choose between competing investments
- Domain-specific measures may be more effective in some cases

Selected references

- Simon HA 1955. “A behavioral model of rational choice”. Q. J. Econ. 69(1):99.
- Gigerenzer G, Goldstein DG 1996. “Reasoning the fast and frugal way: models of bounded rationality”. Psychol. Rev. 103(4):650–669.
- Colyvan M 2008. “Is probability the only coherent approach to uncertainty?” Risk Anal. 28(3):645–652
- French S 2022. “Axiomatizing the bayesian paradigm in parallel small worlds “ Operations Research 70 (3), 1342-1358
- Gutfraind A 2023. “Risk-Reducing Design and Operations Toolkit: 90 Strategies for Managing Risk and Uncertainty in Decision Problems.” *arXiv* http://arxiv.org/abs/2309.03133
- Gutfraind A 2023. "On solving decision and risk management problems subject to uncertainty." arXiv http://arxiv.org/abs/2301.10244 .

Conclusions

- There exists a cross-disciplinary toolkit for risk reduction
- Strategy for AI: use proven RDOT strategies + domain-specific controls
- Top 3: Safety culture, Robustness, Multi-layer defense

Thanks!

Alexander Gutfraind, agutfraind@luc.edu + Vicki Bier, bier@engr.wisc.edu

https://github.com/sashagutfraind/uncertainty_strategies