Trustworthy AI
Gaps and Challenges

Vladik Kreinovich
[1]Department of Computer Science University of Texas at El Paso
El Paso, Texas 79968, USA, vladik@utep.edu

1. Trustworthy Beyond Explainable

- Everyone is talking about Explainable AI (XAI), but trustworthiness is broader than that – and different.
- This difference is easy to explain on the example of people.
- Con men are actually the ones who can explain everything clearly and smoothly.
- And this explainability is a good indicator of a con.
- People whom we trust often cannot explain why they give this advice.
- But we followed their often counterintuitive advice in the past, and it worked.
- It is not possible to directly build trust into a system.
- The future AI system must earn our trust by giving us good advice.
- Also, with trust: one bad step, trust is gone.
- Similarly, a trustworthy AI system must be infallible.

2. Trustworthiness, Privacy, and Potential Conflicts of Interest

- Do I trust Google? No, whatever I ask Google, it keeps it in its records and uses is to help others – and the company.
- Do I trust ChatGPT-type systems? No, if I ask its advice on how to defeat an adversary A:
  - it will store it in its memory, and
  - it may use it to help A when A asks ChatGPT for help.
- When I hire a lawyer to help me, this lawyer is not using my information to hurt my interests.
- We need – really or virtually – individual AI assistants.
- Organizations and groups may also need such dedicated assistants.
- Problem with a current AI assistant? we need a new one.
- In case of conflict between two folks, AI assistants should negotiate.
- These are all new tasks, we need to get prepared.