Dependability Remit Narrow vs Broad

Roy Maxion

Dependable Systems Laboratory
Computer Science / Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213
Email: maxion@cmu.edu

24 June 2023
IFIP 10.4 Invited Talk
Arcos de Valdevez, Portugal

Please do not distribute.

First … thank you to …

- IFIP Working Group 10.4
  - It's an honor to be here.
- Our sponsors…
  - National Science Foundation
  - Computer Emergency Response Team (CERT)
  - Center for Statistics & Applications in Forensic Science
- Collaborators…
  - David Banks (Duke)
  - Vrishab Commuri (UMD)
  - David Eckhardt (CMU)
  - Wangzi He (Microsoft)
  - Abby Martin (ISU/CSAFE)
  - Harvey Vrsalovic (Intel)

Today's message

- Dependability "remit"
  - The task or area of activity officially assigned to an individual or organization.
- Narrow or broad?
  - Narrow: our history, w/ some exceptions
  - Broad: our future (?)
- What will enable the group to thrive?
  - Are we thriving now?
  - What is our mission?
- Today: explore alternatives
  - Some examples
- Tomorrow: thrive or die?
  - What do we, as a group, want?

Brief background

- FTCS-1984: Maxion & Morgan. "The Application of Artificial Intelligence Techniques to Reliable and Fault-Tolerant Computing."
- Lived through the first wave of AI, and then the [first] AI winter; now, in 2023 … déjà vu.
- 10.4 workshop chair
  - 2004, Siena: HCI and Dependability
  - 2009, Obidos: Experimental Methods
  - 2015, Bristol: Science of Cyber-Security
- DSN-19
  - Test of time award (biometric authentication)

Start with AI: large language models

- ChatGPT
  - On-line contradiction; partly truth & partly fiction
  - But which parts? And how do you know?
- Examples
  - Obituaries
  - Bibliographies
  - London bus numbers
    - Traditional: ask your travel agent what the stops are

I don't need no stinkin' travel agent

- ChatGPT:
  - **Q:** Which TfL buses stop at both Bolton's Lane and Hayes & Harlington, outbound from Heathrow Central?
  - **A:** There is no TfL bus that stops at both Bolton's Lane and Hayes & Harlington outbound from Heathrow Central.
  - **FACT:** 278 and N140 each stop at both locations.
- Conclusion: ChatGPT is not reliable. Period. Now what?
- This is AI-generated mis-information.
- What is the risk of blind trust; how do we handle that?

AI – the future

10.4 Retrospective remit

Narrow	Broad

▪ Primary focus:
  ▪ Traditional FT/reliability
  ▪ Hardware
  ▪ Software
▪ Nothing wrong w/ this
▪ But it's limiting
▪ Risk of calcification and demise
▪ What will the future look like?	

10.4 Prospective remit

Narrow	Broad
	▪ Extend our remit ▪ Be a dependability resource for everyone ▪ Let's explore

A lot of "stuff" happens

- 10.4 is essentially in the 'stuff happens' business.
- We've developed ways of thinking about failure.
- ..... And there's a ton of 'stuff' out there.
  - Conference registration page
    - Old/new user? Can't recall using this before, so I must be new.
      - Fill out entire form; then "you already have an account."
      - Then ... recover old password
      - Then ... require creation of new password
      - Then ... gives me a new PayPal account (w/o asking)
      - Then ... daily email from PayPal to finish setting up my new account
  - Hotel internet log-on; interchange name/rm #
  - TSA rejects Colorado IDs (19 June 2022)
  - RUS: Surprise
  - Many, many other examples
- Why is there so much 'stuff'?
- Lots of people get hurt.
- And what gets done about it?
  - Nothing. Why is that?

Consider this

- 1983: Ironies of Automation, Automatica, Bainbridge.
  - Automation causes human skill-set erosion.
- 2004: Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography, Academic Radiology, Alberdi, Povykalo, Strigini, Ayton.
- 2018: Ironies of Automation: Still Unresolved After All These Years, IEEE Transactions on HCI, Strauch.
- Something's wrong here.
- Can 10.4 address it?

Dependable HCI

- There are so many user interfaces that don't perform reliably.
  - Library on-line catalogs (99% reliable)
  - Amazon
  - Visa services
  - Airlines
- Not clear what to do.
- Not fault tolerant.
- Often result in unrecoverable errors.
- Can 10.4 contribute to a solution?

Mess milieu

- Bank-loan denials
- Dependable voting
- Bail decisions
- Parole decisions
- Face recognition

- None of these things work dependably
- People are being hurt every day
- Some outcomes are tragic

- Can 10.4 help?

Undependable science

- Vast public mistrust in science
- NYT, 19 June 2023: The Science of What We Eat Is Failing Us
- The World Health Organization recently advised people to avoid using artificial sweeteners for weight loss or to reduce their risk of health issues like heart disease and diabetes. This was based on the agency's review of available research on artificial sweeteners to date.
- Unfortunately, people cannot be confident in those findings. That's because existing studies on artificial sweeteners are plagued by methodological problems.
- Can 10.4 help with issues of experimental methodology?
  - Teach 10-minute lessons, 3-4 slides?

Judicial system (i)

- 2011 trial (US vs Railey)
- Charge: attempted production of child porn
  - We have the trial transcripts
- Did Railey take the photos, or not?
- FBI testified that he did
- Used FINDCamera: FBI software based on 2009 paper (Goljan et al, SPIE)
  - Relied on photo response non-uniformity (PRNU)
  - Error rate: 1 in a million
- Data never made available; now lost
- Cannot be reproduced.

Judicial system (ii)

- Test data … scraped from Flickr
  - No protocol
- Used auto-exposed images
- But Railey's pictures may have been taken at night
  - too dark?
  - too light (flash)?
- Our group tested over-exposed and under-exposed images
  - Error rate: 1 in 200
  - The jury never saw this (or any other testing)
- Outcome 1: 50 years in prison
- Outcome 2: precedent was established, so all future cases are permitted to rely on FINDCamera
- Can 10.4 live with that? Can 10.4 help?

AI/ML systems are decision systems

- Decisions are informed by [labeled] data.
- What if the data are wrong?
- How would you know?
- How would you test the data integrity?

Undependable data

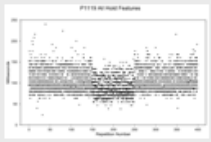
- Examples:
  - Railey FINDCamera case
  - A million images
  - Detect auto/under/over-exposed images
- Keystroke dynamics
  - 1% bad data can change decisions by 20 points
  - Hundreds of thousands of typing records

Automate inspection of a million images

Auto-exposed? Under-exposed? Over-exposed?

And how do you validate the result?

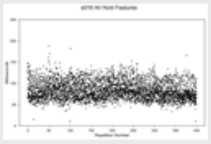
Automate inspection of n000 records



Bad data

But you have no idea where to start; no up-front clue for separating good/bad data.

Note: most data are bad.



Good data

The quality of research in my own field

- Problems in keystroke biometrics
  - Performance comparisons on disparate data sets
  - No power analysis; insufficient subjects/data
  - No data-gathering protocol
  - No common data sets (except perhaps CMU)
  - No common algorithms
    - New alg tested on new data, then compared to other algs on other data; a farce
  - No standard terminology; reinvented concepts
  - Noisy data
  - Bad data (e.g., USB-affected)
  - Poor measurement, poor statistics
  - Methodologically unsound … in the extreme
- Can 10.4 help?

Safety / Assurance cases

- A reasoned and compelling argument, supported by a body of evidence, that a system, service or organization will operate as intended for a defined application in a defined environment.
- Much needed, and much to be done.
- Can 10.4 help?

Now what? Narrow vs broad

Narrow

- Stay the same
- Minimal need for new people
- Minimal effort
- Risk of calcification
- Possible irrelevance

Broad

- Extend remit to new topics
- Recruit new members with relevant expertise and interest and leadership ambitions
- Consider periodic publications sponsored by the group; make 10.4 a force in broadly perceived dependability.

The ideal 10.4 meeting? Ideas …

- Topics that encourage wide engagement
- Technical presentations that
  - are informative to a non-specialist
  - are succinct, clear overviews of key emergent issues
- No NDA restrictions
- Guests who can speak tech
- Everyone learns something
- New colleagues; new members
- Opportunities open up for collaboration
- Cross pollination across communities
- Avoid arcane details (unless absolutely necessary)
- Open dialog, not restricted to end of session; speaker manages time
- Wide-ranging participation in discussions
  - at end of session
  - at meals and breaks
- Establish 10.4 imprimatur
- Publish 10.4 outcomes, decisions, opinions, recommendations
- Expand publishing venues
  - blog/opinion pieces may be more persuasive to current audiences

What is your ideal 10.4 meeting?

- If you knew in advance that …
  - x would [not] happen in a meeting
    - I would [not] attend
- What are your Xs?
- What are your criteria?

Conclusion

- Risk of 10.4 stagnation, irrelevance
- Examples of new undertakings
- Criteria for "good" meeting
  - Always in the eye of the beholder
  - But needs broad agreement among members
- Thank you … now let the discussion begin.

- End – End – End – End – End – End -