Origins of RE for AI: A Hot Topic as Viewed by an RE Alter Kaker

Daniel M. Berry
University of Waterloo

Abstract

I explore how not-very-successful attempts to fit requirements for artificial intelligences (AIs) and learned machines (LMs) into the traditional RE mold led to a rethinking about RE for AI. I talk about some implications of the rethinking.

The full set of slides for this talk is at

cs.uwaterloo.ca/~dberry/FTP_SITE/tech.reports/RE4AIoriginsSlides.pdf

Vocabulary

AI **artificial intelligence**

ML **machine learning**

LM **learned machine**

These slides use “AI” to mean “AI or LM” to save slide space and speaker’s breath.

David Parnas's Concerns

Back in 2019, I had seen slides from Dave Parnas expressing very negative concerns about AI in his inimitable way.

Unpredictable Behaviors of AIs

He shows several examples of very bad side effects arising from the unpredictable behavior of AIs.

I won’t show all of his slides.

Dave Parnas's Letter to CACM

Instead I will quote a letter that Dave wrote to the editor of *CACM* that appears in *CACM* August 2019 (62:8) on Page 9:

"[The dangers are] not limited to neural networks or machine learning technology. [These] dangers … exist whenever a program's precise behavior is not known to its developers. …

Dave Parnas's Letter, Cont'd

… I have heard neural network researchers say, with apparent pride, that devices they have built sometimes surprise them. A good engineer would feel shame not pride. In safety-critical applications, it is the obligation of the developers to know exactly what their product will do in all possible circumstances. Sadly, we build systems so complex and badly structured that this is rarely the case."

— David Lorge Parnas

Dave Parnas's Complaint

Dave's complaining that AIs are not behaving logically, in predictable manners, and are stochastic in their behavior …

like the humans that they are replacing.

We are giving AIs powers to do things *faster* than humans can and at a *bigger* scale, without any way to guarantee that they will do so *better*, *more reliably*, and *more predictably* than humans can.

Very Frightening! ☹

This is the stuff of technology-based horror movies!

E.g., Star Trek's *Doomsday Machine* or V'ger

An AI could have a *more widespread* catastrophic effect *faster* than we humans are able to notice it and stop it.

A Graduate Seminar I Taught

Fast forward to 2019, to my "Advanced Topics in RE" graduate seminar at UW

I asked the AI PhD students in the class to prepare their paper and talk on RE for AI, …

hoping to learn what AI people consider to be a specification for an AI.

Wotta Disappointment!

They repeated all the usual RE parenthoods:

correctness
completeness
consistency
robustness
reliability

Yeccchhhhh!!!!! ☹

A Conversation

between me (D) and one of the students (S), who shall remain nameless to protect his reputation:

At the end of the talk:

D: OK.. but what *is* correctness?

S: (shrugs)

D: I mean, how do you *know* that the AI you have produced is correct?

A Conversation, Cont'd

S: You don't! It's probably *not*!

D: That's horse s--t! I have seen AIers continuing to revise their AI until something was true. I would hope that that the something is correctness. So what *is* that something?

S: Ah! They keep at it until the recall is high enough and the precision is low enough.

D: Ah! So how do you know that the recall is high enough and the precision is low enough?

A Conversation, Cont'd

S: We guess! We just feel it!

D: Ewwwww!

Want to Use the RE Ref. Model

Dave and *I* desperately want an AI to behave like SW in the RE Reference Model, the (Zave–Jackson Validation Formula) ZJVF …

to be able to validate that an AI is behaving as expected, i.e., as specified.

Reference model

- Thus, if we enlarge our model to include domain knowledge, then the following relationship must hold:

$$\mathbf{D, S \vdash R}$$

  - D is domain knowledge
  - S is the specification
  - R is the requirements

- The specification describes the behaviour of a system that realizes the requirements.
- The domain assumptions are needed to argue that any system that meets the specification (and that manipulates the interface phenomena) will satisfy the original requirements.

Hidden: Traffic light example

- D = drivers behave legally and cars function correctly
- S = spec of traffic light that guarantees that perpendicular directions do not show green at same time
- R = perpendicular traffic does not collide

Problem: make D unnecessary, steel walls pop up on red, light controls cars by wireless

Uncertainty in "D, S ⊢ R"

- The formula D, S ⊢ R tries to be formal in the sense of describing what happens completely.
- One would expect computers and software and their combination to be formal in this sense.
- But, the real world intervenes to make this formula only a guideline and not an accurate, precise model.

Uncertainty in "D, S ⊢ R"

- The formula D, S ⊢ R tries to be formal in the sense of describing what happens completely.
- But, as we have seen, it cannot be completely formal because at least D and R have to describe the real world, which is not formal

What does this do to the hope of formally modeling computer systems?

Molecular Software

- Molecular SW, e.g., DNA, RNA, Proteins, Catalysts
- Molecules designed specifically to achieve a desired effect
- Molecule is shown empirically to behave as specified in S, with 99.95% certainty
- In this case, in D, S $\vdash$ R, also S is informal!

AI's Behavior is Stochastic

Since the behavior of an AI is stochastic, like molecular SW, the truth of S is empirical, not logical.

So now validation of an AI has three empirical truths instead of just two, and logic plays almost no part.

Brooks's "No Silver Bullet"

Harel's "Biting the Silver Bullet"

Biting the AI Bullet

Instead of Parnas's despair and unrealistic hope of prohibiting AIs altogether …

or at least making them non-stochastic or logical in their behaviors,

we need to bite the AI bullet.

Biting, Cont'd

Change the nature of a specification of an AI to take into account the stochastic behavior and give empirical measures of acceptable behaviors …

Biting, Cont'd

so that validation becomes akin to empirically proving the hypothesis

"The AI behaves as specified."

and we give confidence intervals or *p* values to the claims of how well the measures are matched …

***and* we make engineering judgements for close calls.**

Leading to REFSQ 2022 Paper

This realization led to the paper I presented at REFSQ 2022:

“RE for AI: What is an RS for an AI?”

Slides from REFSQ 2022 Talk

RE for AI: What is an RS for an AI?

Daniel M. Berry
University of Waterloo

Main Insight of REFSQ Paper

The main insight of my REFSQ'22 paper is that a specification for an AI for a hairy task consists of

1. **a set of measures used for evaluation,**
2. **criteria that the measures must satisfy, and**
3. **other data about the context of the use of the AI, including the RW data that teaches an LM.**

Set of Measures

The set of measures used for evaluation measures *correctness* in some sense and is usually calculated from a confusion matrix, e.g.,

- **recall and precision,**
- **sensitivity and specificity,**
- **F-measure**
- **accuracy**

Criteria For Measures

The criteria that the measures must satisfy …

help show that the AI …

can be considered as …

mimicking or doing better than a human doing the same task.

Criteria, Cont'd

These criteria will usually include …

the values of the measures that humans *actually* achieve …

when doing the same task.

Other Data

The other data are data about the context of the use of the AI that …

- **allow engineering tradeoffs to help the AI meet the criteria and**

- **decide borderline cases.**