Validation Challenges in Artificial Intelligence

Douglas A. Samuelson
InfoLogix, Inc.
Annandale, Virginia
infologix1@aol.com

Society of Decision Professionals,
April 18, 2024

The Uncanny Valley

- People distrust and dislike AI/ ML systems that are “too close” to human behavior but distinguishable
- There seem to be parallels in how humans and other organizations are perceived, too – bosses want to multiply subordinates, not rivals

The Trust Problem

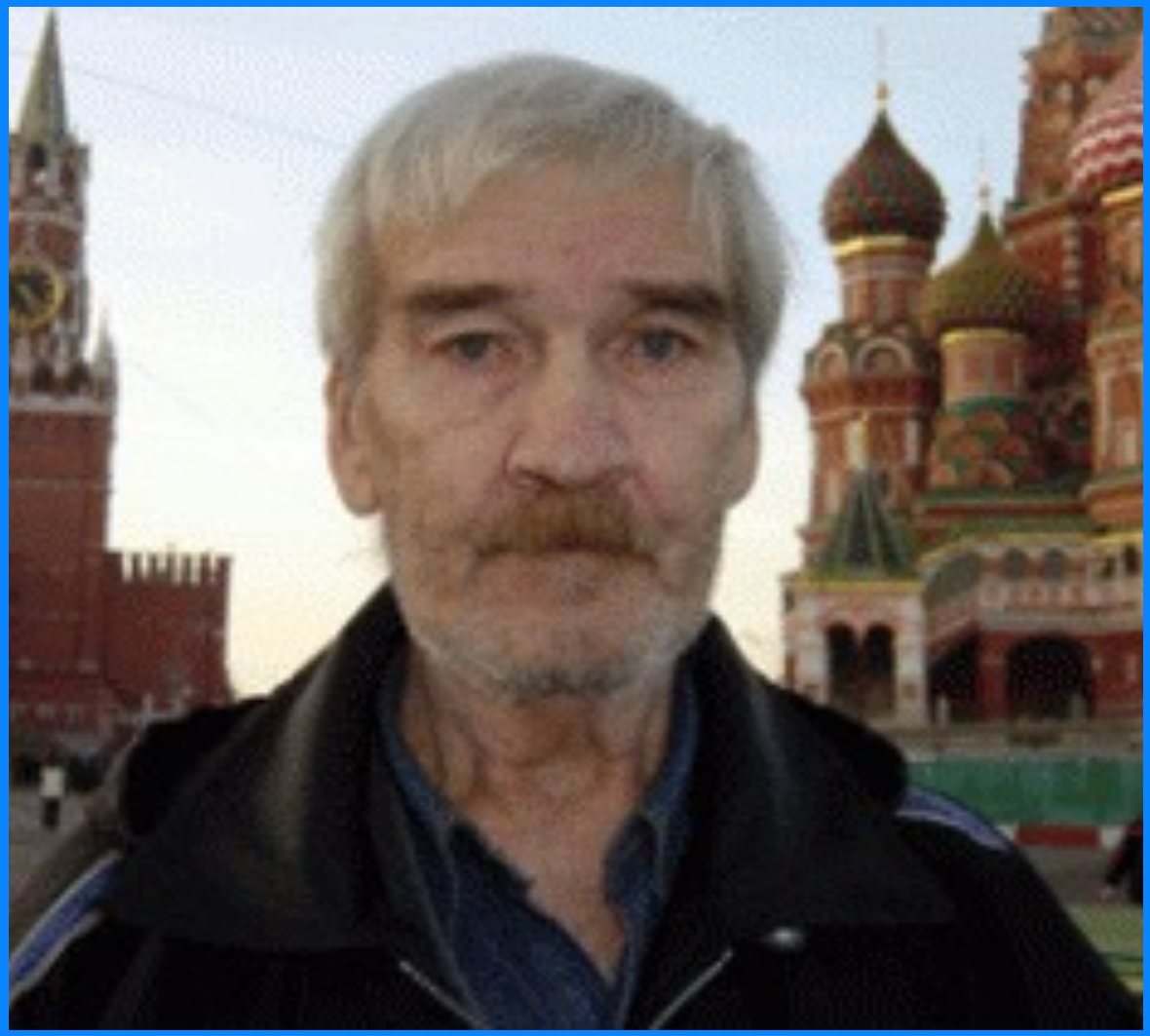
- So how do we know which systems to trust?
- For what decisions?
- Would you follow a kill recommendation from this system or subordinate?

A Portentous Night

- September 26, 1983: newly upgraded Soviet satellite systems reported five US ICBMs launched toward USSR
- LTC Stanislav Yefgrafevich Petrov chose not to report upward – system output "just didn't feel right" (they had wargamed US attack patterns)
- Commended for being right, sidelined for disobeying reporting orders

Let this sink in…

Civilization still exists right now because this guy didn't follow his orders!



When Should One Resist Orders?

Every US military person is taught to resist illegitimate orders. (UCMJ)

How to determine legitimacy of orders is not clear at all. It's "gut feel."

How do we teach "gut feel" to a machine? How does a machine learn context?

Important because:
…the coming Swarm Storm

- Developing capabilities for swarms of UAVs to combat each other and seek additional targets
- Action too fast and too complex to be managed or checked in real time by humans
- What limits can / should the controlling computers have?

Controlling the Berserkers

Berserker: a wild, single-minded warrior

As with human subordinates, commanders need to have early indicators of unacceptable behavior

This means having metrics of what is expected to happen, and of what would be a likely violation of expectations

Managing the Berserkers

If you can't control individual actions, you need to establish and enforce codes and protocols of conduct

At a more general level, you create a culture of expectations and goals

Progressively weed out the ones who don't fit the culture

This is difficult enough with humans – even harder with machines, which lack context

Identifying the Berserkers

Space Shuttle: multiple computers vote on what to do, outlier gets reexamined and may eventually get shut down

Develop metrics of leading indicators of cognitive entities that may be headed for trouble: does little things now that are associated with big misdeeds in crises

Identifying the Trustworthy

One common task for AI systems is to identify another entity's logical processes. We can test how well our system does this on processes with known logic.

How?

Additional challenges

AIs require LOTS of data to learn

AIs cannot infer context

AIs generally cannot infer means of inference

(Can't figure out how to find the 16th most populous city in the US, for example)

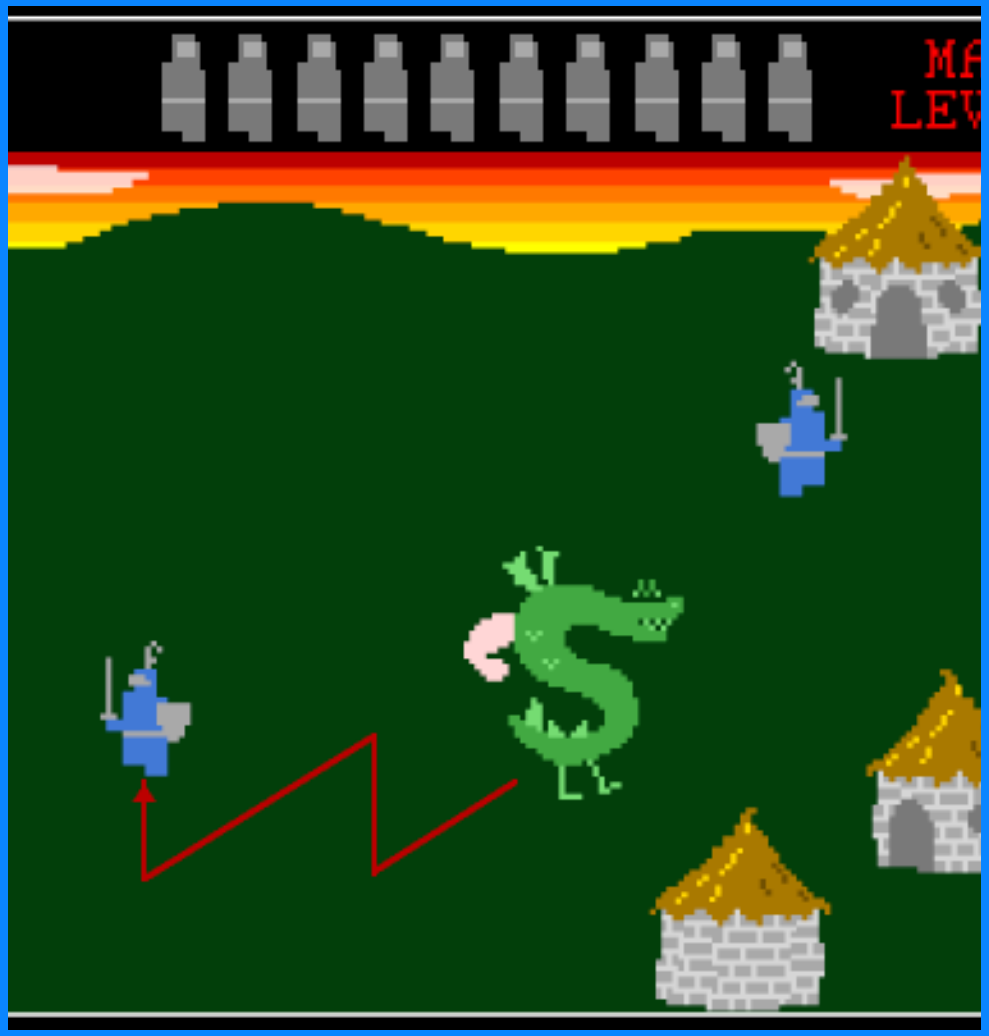
Often can't explain how they reached a conclusion

The Trogdor Test!

"Trogdor" is an online game (homestarrunner.com) in which a dragon moves among villages, stomping on peasants. You move him with the arrow keys.

The Trogdor Test!

He is pursued by two knights whose movements are random at first, but will home on him if he avoids them long enough. Moving between them resets them.



The Trogdor Test

The logic of the knights' movements is programmed into the computer. Any AI worthy of the name should be able, by playing the game, to deduce the knights' control logic exactly. And we know whether it has done so correctly.

In general, how well an AI finds known solutions in tests is a trust metric.

But how many "Adversarial reasoning deducing" AIs do this?

The Chinese Room

"Searle's Chinese Room" is a well-known thought experiment. We have two Chinese translators in opaque, sound-proof rooms. We hand them a piece of writing in Chinese and they pass out an English translation. If we can't tell the difference, and one of them happens to be just a guy with a dictionary, how expert is he in Chinese?

The Chinese Room Fallacy

Searle claimed that this demonstrated that all the intelligence in an AI system is really in the human-created software, not the machine.

But the setup is contrived and forced!

Searle excludes the obvious idea of asking each translator to critique the other's translation: a mutual challenge test.

Why Compare Translations?

Good translation may require understanding idiom.

For example, in 2009, Sharon Stone posed topless for Paris Match.

The headline read,

“J’ai 50 ans – et *alors!*”

Why Compare Translations?

Some US writers with dictionaries (and no knowledge of idiomatic French) rendered this as “I’m 50 years old – and then some!”

The actual meaning of “alors” in this usage is either “So what!” or “Look here!”

Comparing Translations – and Even Then….

Inflection matters. Ironic use matters.

"I just got a flat tire and it's pouring rain.

Isn't my life just *wonderful!*"

Or: "Stalin: You were right. I was wrong. I should apologize."

Properly rendered as, "You were right?! I was wrong??!? *I* should *apologize*??!?"

And speaking of context….

We want our AI to learn what success looks like.

So let's ingest enough data to wargame the Japanese attack on Pearl Harbor, 1941.

Was that mission a success?

Tactically, yes. Complete surprise, six battleships put out of action.

And speaking of context….

But strategically, not good. They missed the submarine refueling facilities and the carriers were out on maneuvers. Four of the battleships were refloated and returned to service within a year.

Yamamoto: “I fear that all we have done is awaken a sleeping giant and fill him with terrible resolve.”

Sharon Stone Revisited

Most browsers have routines to flag potentially mature-audiences or outright offensive content. You could think of these as simple AI.

Try searching (on a non-Government computer) “Sharon Stone topless Paris Match” and see what you get.

I got mostly censored versions of the Paris Match photo, but I also got a few full frontals, with no warning at all.

And check these videos….

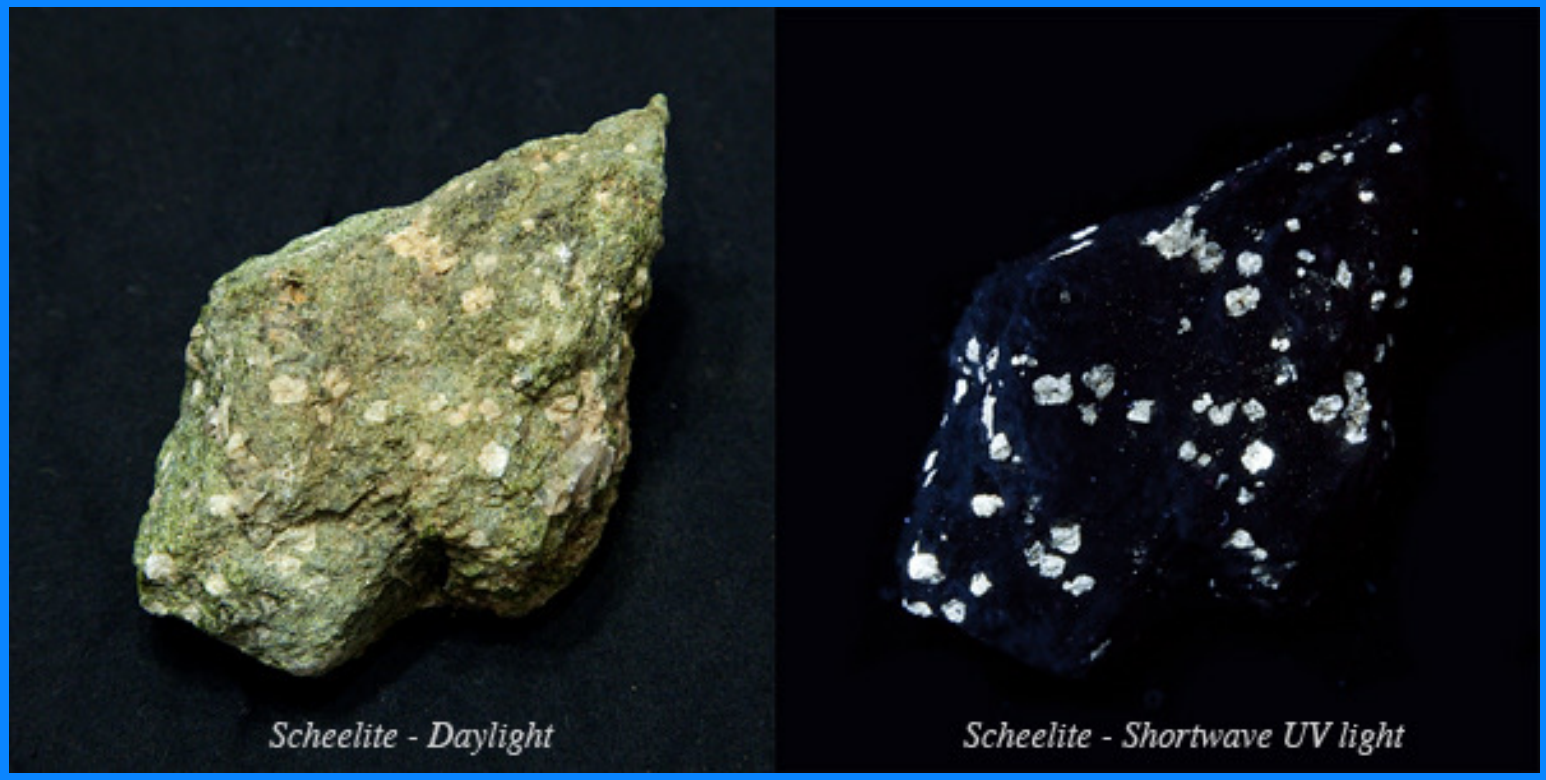
The same screening routines will cheerfully allow you or your kids, without warning or restriction, to view the “Baptism Murders” scene from *The Godfather*, in which eight people get shot to death on screen in less than two minutes. The horrifying machine gunning of Sonny Corleone, earlier in the movie, is also readily available. So tell me again how easy it is to teach standards to a machine?

How reliable are observations?

We want to validate AIs, simulations, wargames, etc. against highly reliable direct observations. Example:

How reliable are observations?

So what? Well, let's change the light….



How reliable are observations?

All observations require using some medium – light, electron beams, sound waves, ….

And the medium contains inherent uncertainty.

Generalization of the Scott Effect: distant faint objects may be defects in the optics!

Therefore, **all observations are somewhat uncertain.** (Sorry, Albert!)

So What Can We Do?

Challenge test and/or wargame and use “red team” experts (and non-experts) to develop better understanding of what could be done, what effects it could have, where knowing a little more would change your decision

Express both inputs and outputs in storytelling form, quantify only in meaningful ways

(E.g., how many stories contained this term? Is there a pattern to where they occurred?)

Testing, testing…

Take a lesson from "Dirty Harry" Callahan, in "Magnum Force": A man's got to know his limitations.

This is true of AI / ML systems, as well.
Never trust a system or its developer if the developer claims the system has no limitations.
What that really means is that they never tested it.

Another Lesson Learned the Hard Way

Never enter into a NDA covering internal logic with a software provider.

Khrushchev used to tell about a man who ran through the streets of Moscow shouting, “Khrushchev is a fool.”

He was sentenced to 16 years in Lefertovo Prison: one year for publicly insulting the Party General Secretary, and 15 for revealing a state secret.

The Hard Lesson

If the software fails miserably, everything you can say to defend a trade libel suit for stating that it failed turns out to be revealing confidential information.

There are similar “gotchas” in general in the discussion of classified information. Non-disclosure agreements require great care: limit the scope, the remedies, and the time it remains in effect.

Another Hard Lesson

Theorem: no AI system can infer actions or consequences about which none of the people teaching it know anything, and for which no experience database is available.

Specific example: AI cannot credibly predict the actions of a party with which nobody in the wargame is familiar. At best you get digitized depictions of your guesses.

Yet Another Hard Lesson

Digitizing a conjecture adds precision but not information.

Theorem (the famous one): the more closely a computer model approaches the complexity of reality, the more difficult it becomes to distinguish genuine rare emergent effects from deficiencies in the programming.

Connection to Big Data

Derive as much as possible from narratives: what were players thinking given what they knew?

Same analytical process works for interpreting Big Data outputs, not just results from wargaming – with the same pitfalls

“What is this telling us? Are there patterns? How consistent are the data?”

Keep Thinking of “What If…?”

Just one bad assumption about what “can’t happen” can lead to big trouble!

The Critical Challenge

Thomas Schelling (Nobel laureate, 2005): Nobody can make a list of all the things he never imagined.

Summary and Conclusions

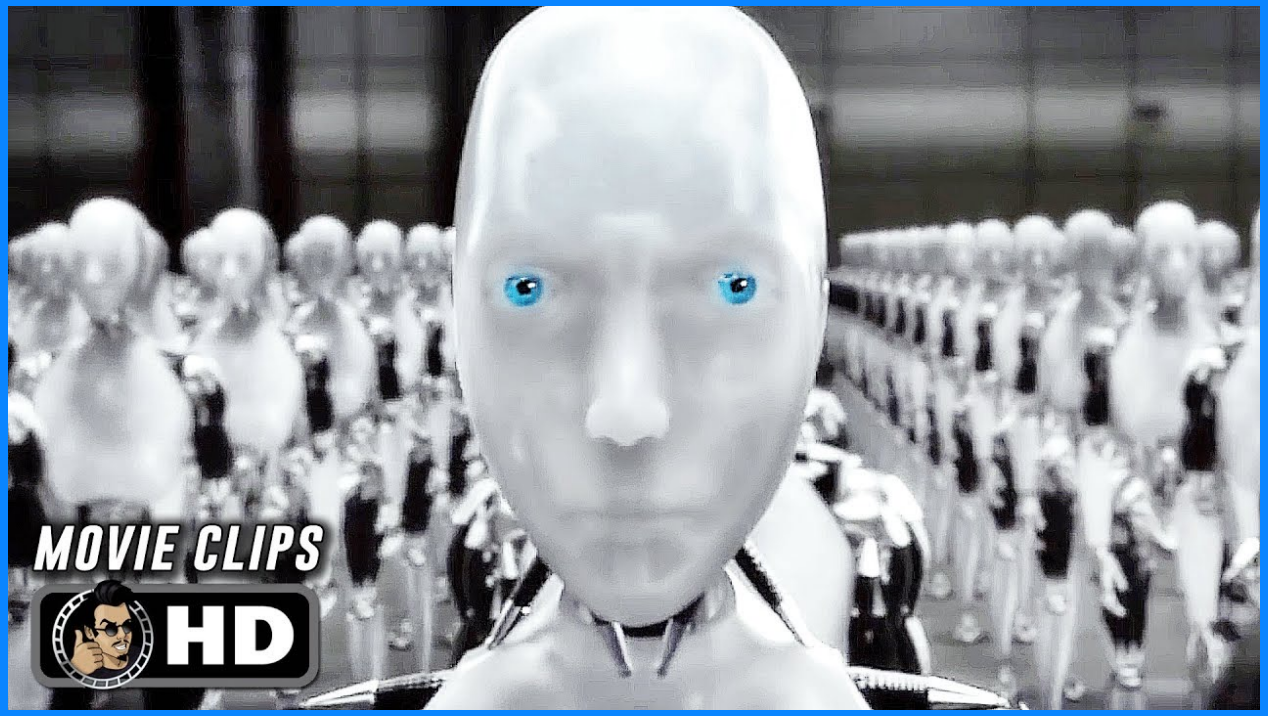
AI and Data Science can and must develop ways to improve and critique each other's analyses: **challenge test!**

We need to develop metrics to help us understand which methods and approaches are trustworthy

We need to pay particular attention to identifying topics of interest we haven't considered yet

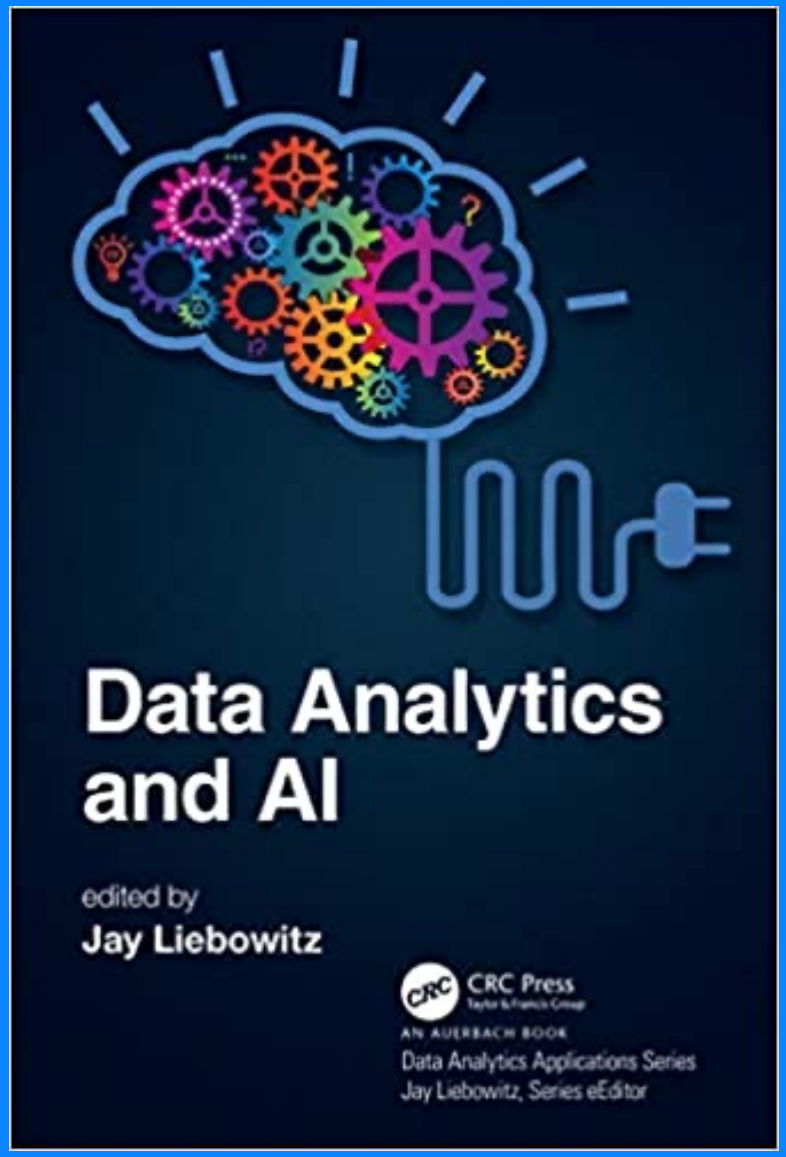
Final Warning

Do you think you can develop AI that can out-think us but will not harm us?



Be careful about what you wish for!

For Further Reading….



Concluding Thoughts

No one of us knows more than all of us.

Diversity outperforms ability, because diversity reduces the number of blind spots.

-- Scott Page

"The greatest threat to our security is inertia in the thinking of those responsible for our security."

– David Ben-Gurion

References

Cole (2004), https://plato.stanford.edu › entries › chinese-room retrieved January 9, 2020.

Samuelson D., (2020) "Measurement Issues in the Uncanny Valley," *Data Analytics and AI,* J. Liebowitz, ed., Auerbach, August

Trogdor: http://homestarrunner.com/trogdor-canvas/index.html , retrieved January 8, 2020.

Uncanny Valley: https://en.wikipedia.org/wiki/Uncanny_valley retrieved January 7, 2020.