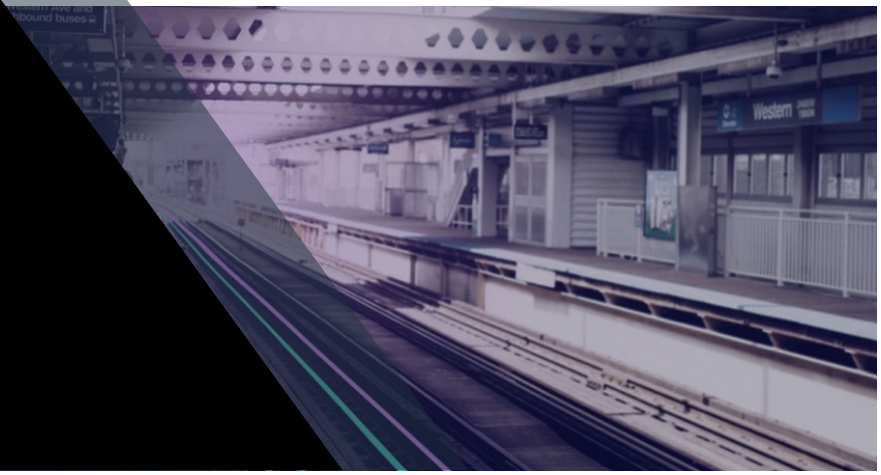
ACL Workshop: The Future of Benchmarking

August 2021

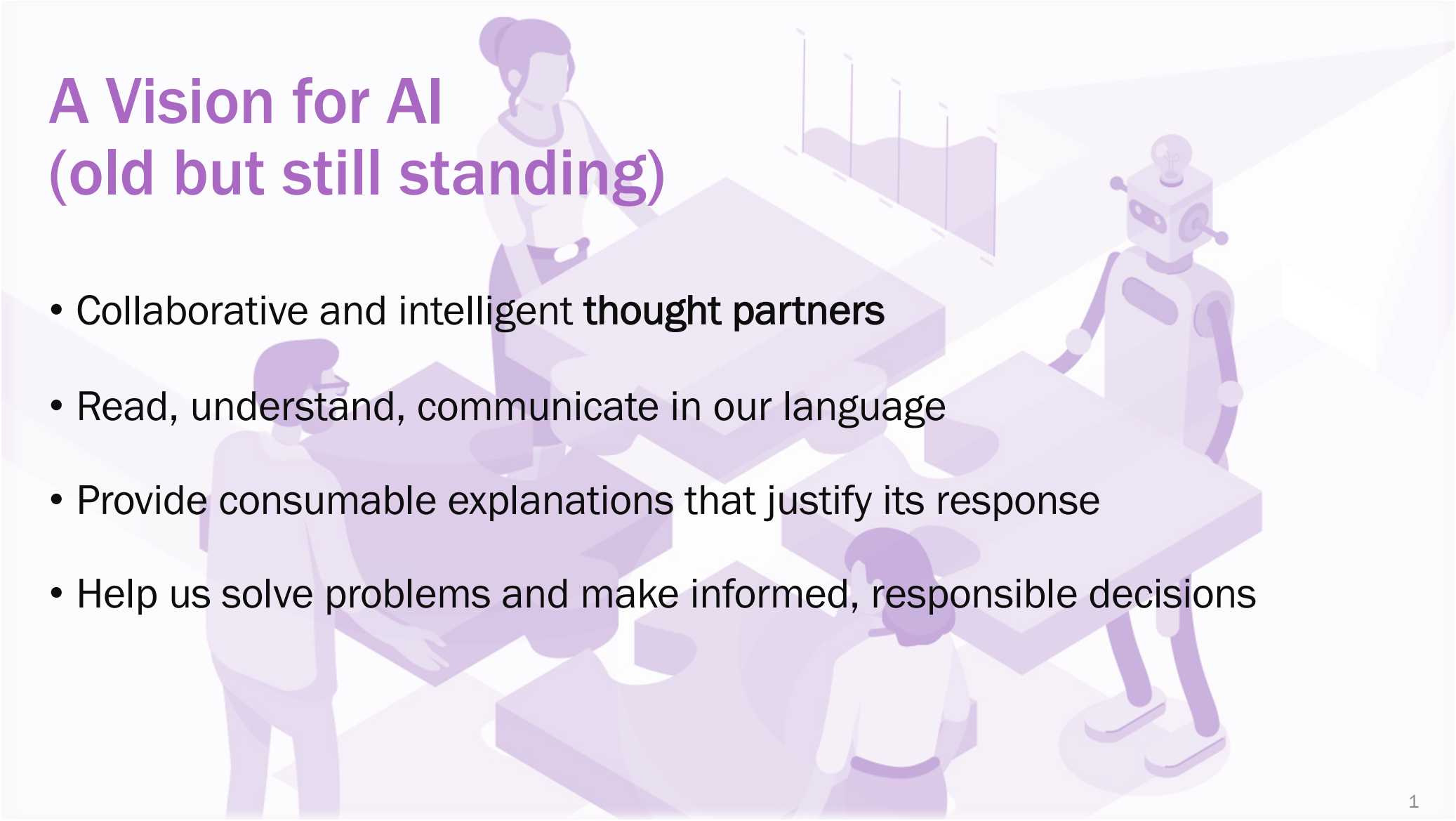
Dave Ferrucci

Measuring Understanding

ELEMENTAL cognition



A Vision for AI (old but still standing)



- Collaborative and intelligent **thought partners**
- Read, understand, communicate in our language
- Provide consumable explanations that justify its response
- Help us solve problems and make informed, responsible decisions

We lack a clear vision for testing that machines understand

- Intentionally solving different problem?
- The problem is too subjective?
- Too expensive to evaluate?
- Too hard?
  - Will make us look bad?
  - Or at least like we are making much less progress?
- Too big and slow for the academic paper mill?

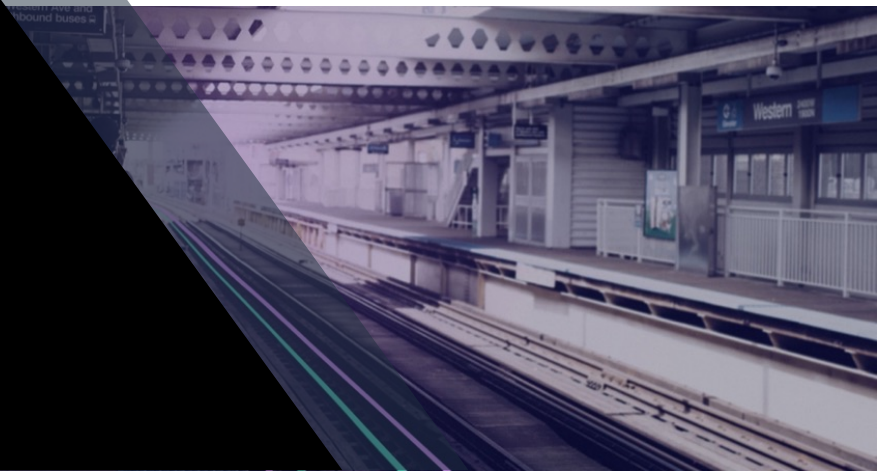
Some Principles to Consider

Ambitious	A clear and ambitious **Vision** Define what we really want. With NO regard for the methods that may or may not work.
Realistic	Set realistic **Expectations** for how hard In order to make the right level of investments develop a sense of how difficult it will be do and measure
Holistic	Think Holistically Creatively decompose to solve but always put it all back together to judge holistic performance. The affect of the UX/UI can can change how we architect the guts of intelligent systems
Valuable	The Benchmark Costs < Expected Value of Success Build tests for the big vision based on expected value If test seems too expensive maybe the vision is not ambitious enough or maybe it is believed impossible.

ELEMENTAL
cognition

How hard is it?

Are we testing the right stuff to find out



Understanding

The interpretation of a particular situation in order to provide, the context, insight and foresight **REQUIRED for effective human decision making.**

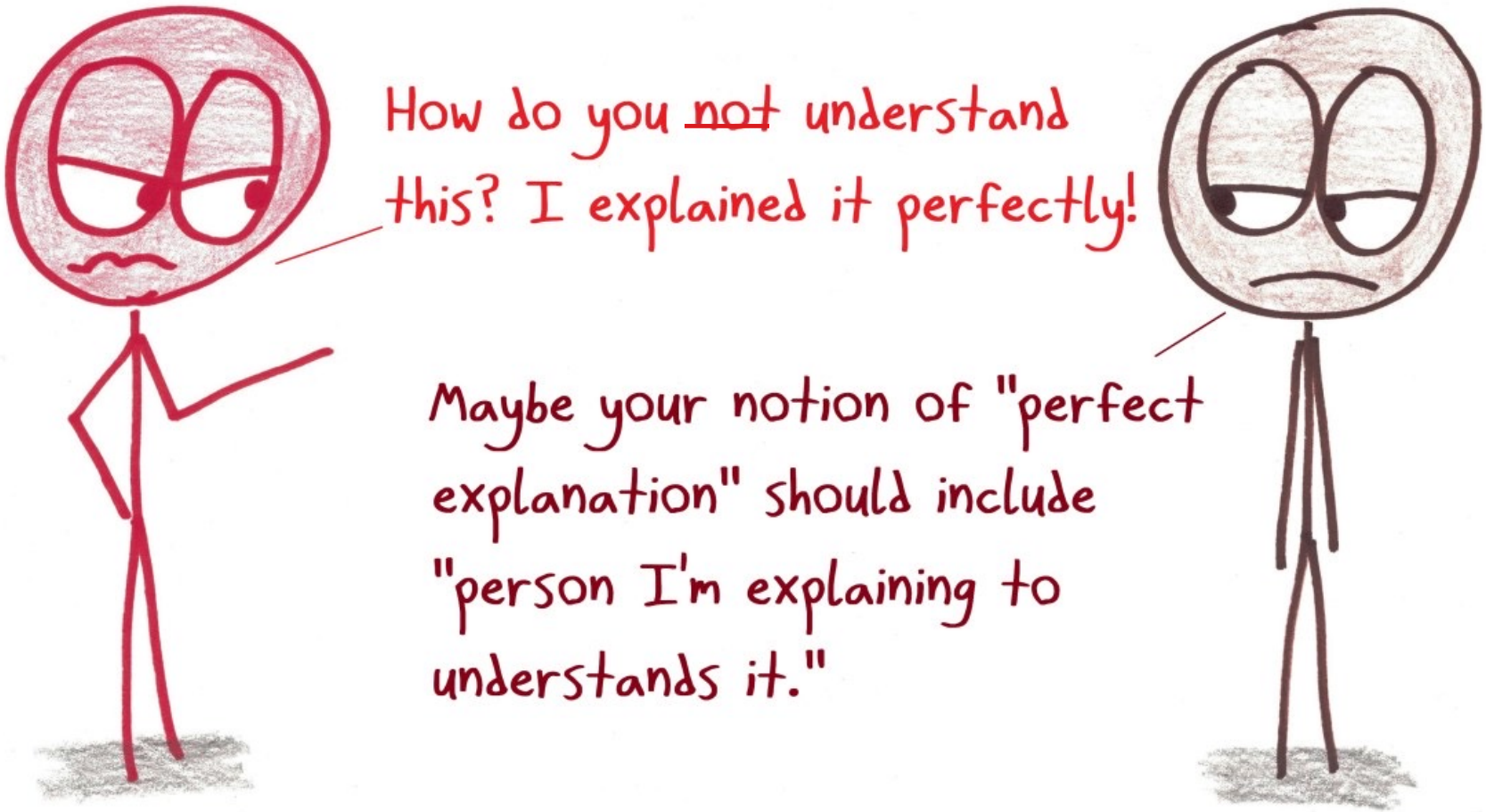
The quality or utility of the *interpretation depends on what decision is being made and even by whom (their prior model).*

Heads up: Its probably hard

We spend enormous and continuous intellectual effort to build and manage shared understanding.

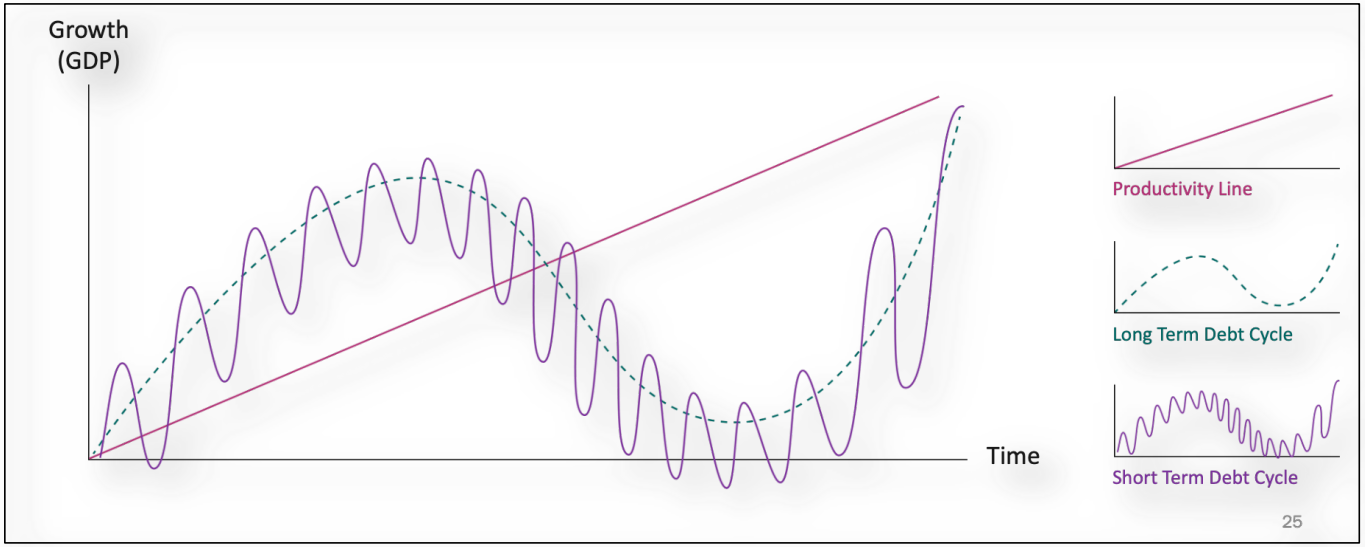
We work hard because it is complex, contextual across multiple dimensions and abstract.

And arguably still fail in many cases.



Contextual: Is the “right” or “better” explanation one that the user can readily understand. If so, the user’s prior becomes an important part of the context.

Are we headed for a recession?



Explanations can involve connecting many variables interactions, long causal chains and layers of knowledge …. or NOT?

Are we headed for a recession?

Which explanation is better? Why?

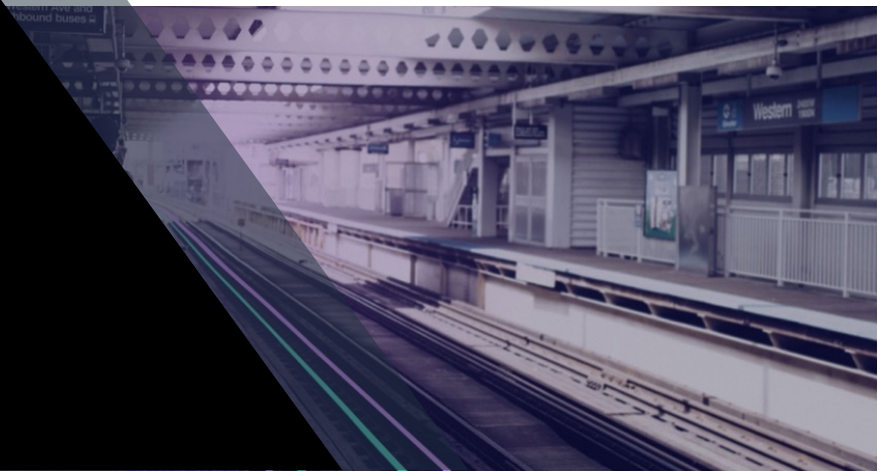
- What level of detail is required for an acceptable understanding.
- Does it have to be causal for a good explanation or understanding

- **Yes**. "…the most reliable forward-looking indicators are now collectively behaving as they did on the cusp of full-blown recessions…"
- **Yes.** "Consumers have extended too much credit to pay for their homes that the housing bubble had made unaffordable. Many, realizing they were over committed, stopped making their payments. The defaults are likely to result in substantial losses for the banks. The degree of leverage [debt] in the system would compound the problem, paralyzing the financial industry. Without access to money in the form of credit, business reduce expenditures, consumers with debt and expectations to make less money, further reduce spending. The drop in demand is likely to trigger a recession.

Adapted from the book: The Signal and the Noise by Nate Silver

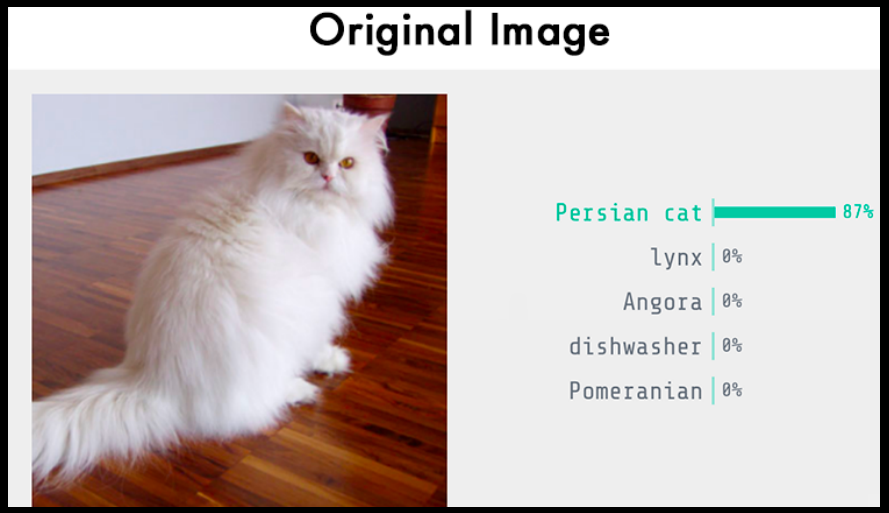
“Telling Errors”

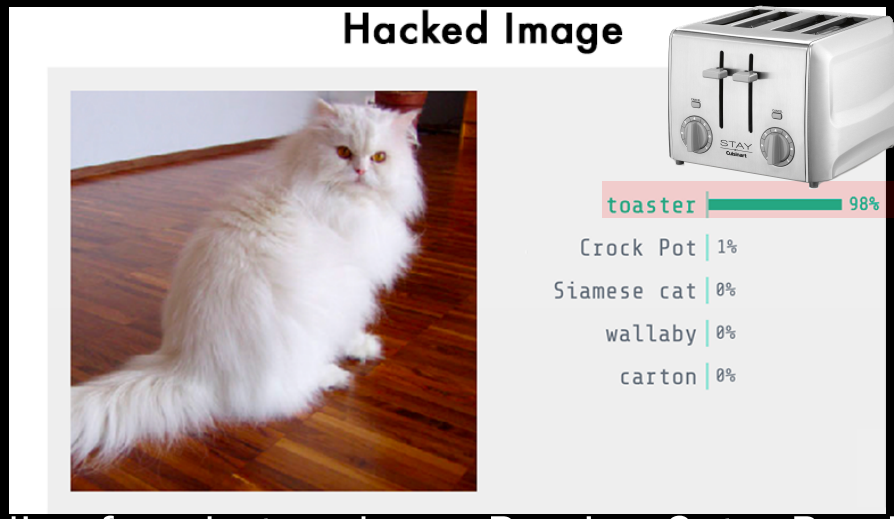
ELEMENTAL
cognition



A Cat or a Toaster?

The Understanding Goal or target seems to really matter in how you judge. Should explicitly include a consumable understanding if that is what is expected result.





This mistake implies a very different understanding for what makes a Persian Cat a Persian Cat ...

But to be fair, the system was trained to classify **NOT** to build a **compatible model that humans can understand. Maybe because the test did not require it.**

How about this one from GPT3?

PASSAGE

The boy approached the turtle. The turtle pulled itself into its shell

QUESTION

Why was the turtle afraid?

ANSWER

because the boy was carrying a large, **sharp knife.**

Good or bad? Should we assume that turtles understand and generate fear from the threat of big knives.

Is there an established prior model our shared understanding that suggests that?

Should GPT3 you get partial credit for this one?

Answer is wrong, but explanation seems to suggest a total lack of understanding.

PASSAGE

John put the sandwich in the lunchbox. He put the lunchbox in the car.

QUESTION

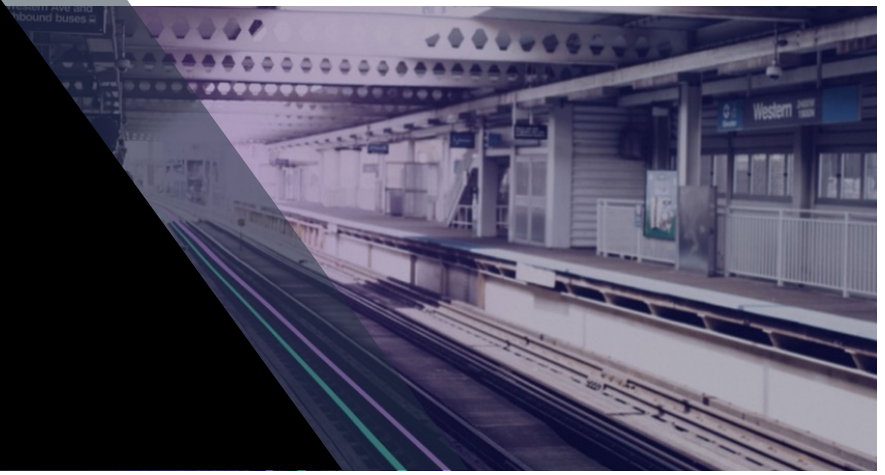
Is the sandwich in the car?

ANSWER

No....The lunchbox is in the car. The sandwich is in the lunchbox.

An initial experiment to raise the bar

ELEMENTAL cognition



Given a “story”…Fill in a template of understanding

Generate and answer these questions in a way people can understand.

SPATIAL

Where are the entities relatively positioned and oriented throughout the text?

TEMPORAL

What events/states occur and with what relative timing?

CAUSAL

How do events/states mechanistically lead to other events/states described or implied by the text?

MOTIVATIONAL

How do agents’ beliefs, desires, goals, and emotions lead to their actions?

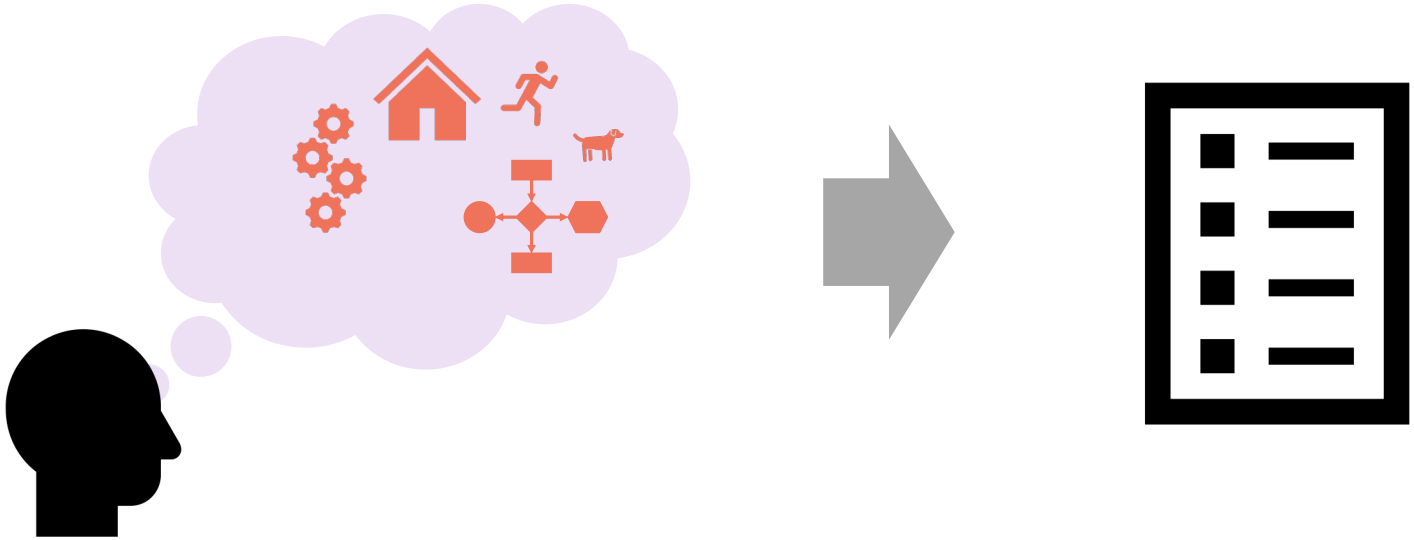
Includes 4 classes of questions reflecting the content human readers attend to, based on research in cognitive science*, to capture and communicate a general “common sense” understanding

* See Graesser et al., 1994; Zwaan et al., 1995

An Evaluation Approach

Have annotators commit their understanding of a story by writing down the ToU for that story.

Provide the **map, time-line, causal graph and the motivations** of all events, agents and state changes in the story. (Even for the "obvious" stuff. Everything!)



An example

excerpt from a “record of understanding”

Sample story fragment:

> ...*One day, it was raining. When Allie arrived, Rover ran out the door. He barked when he felt the rain. He ran right back inside.*

Spatial *(sample entries)*:

- Rover is in the yard from when he **runs out the door** until he **runs inside**.
- Rover is in the house from when he runs inside until the end of the story.

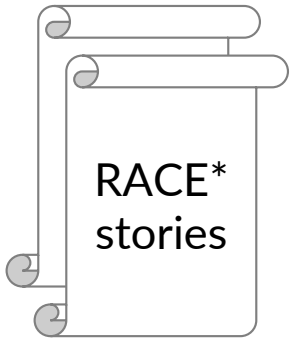
Temporal *(sample entries)*:

- Allie arrives **just before** Rover runs outside.
- It is still raining at the end of the story.

Motivational and Causal *(sample entry)*:

- Rover runs inside, rather than staying put, **because**:
  - If he runs inside, he will be inside, whereas if he does not, he will be outside, because: ...
  - If Rover is inside, he will not get rained on, whereas if he is outside, he will, because:
    - It is raining.
    - When it is raining, things that are outside tend to get rained on, whereas things inside do not.
  - Rover prefer not getting rained on to getting rained on, because: ...

We prefer free-form answers to better test explanations and explicability…
But to test existing systems, we built a small multiple-choice dataset based on our ToU.



Rough RoU

Spatial:
- …

Temporal:
- …

Causal:
- …

Motivational:
- …

Q1a) What happened when Mr. Green and the man drove together?

A. **They came to a small house.**
B. They came to a hotel.
C. They traveled around the country.
D. They stopped several times at the side of the road.

Q1b) How did the man's directions turn out?

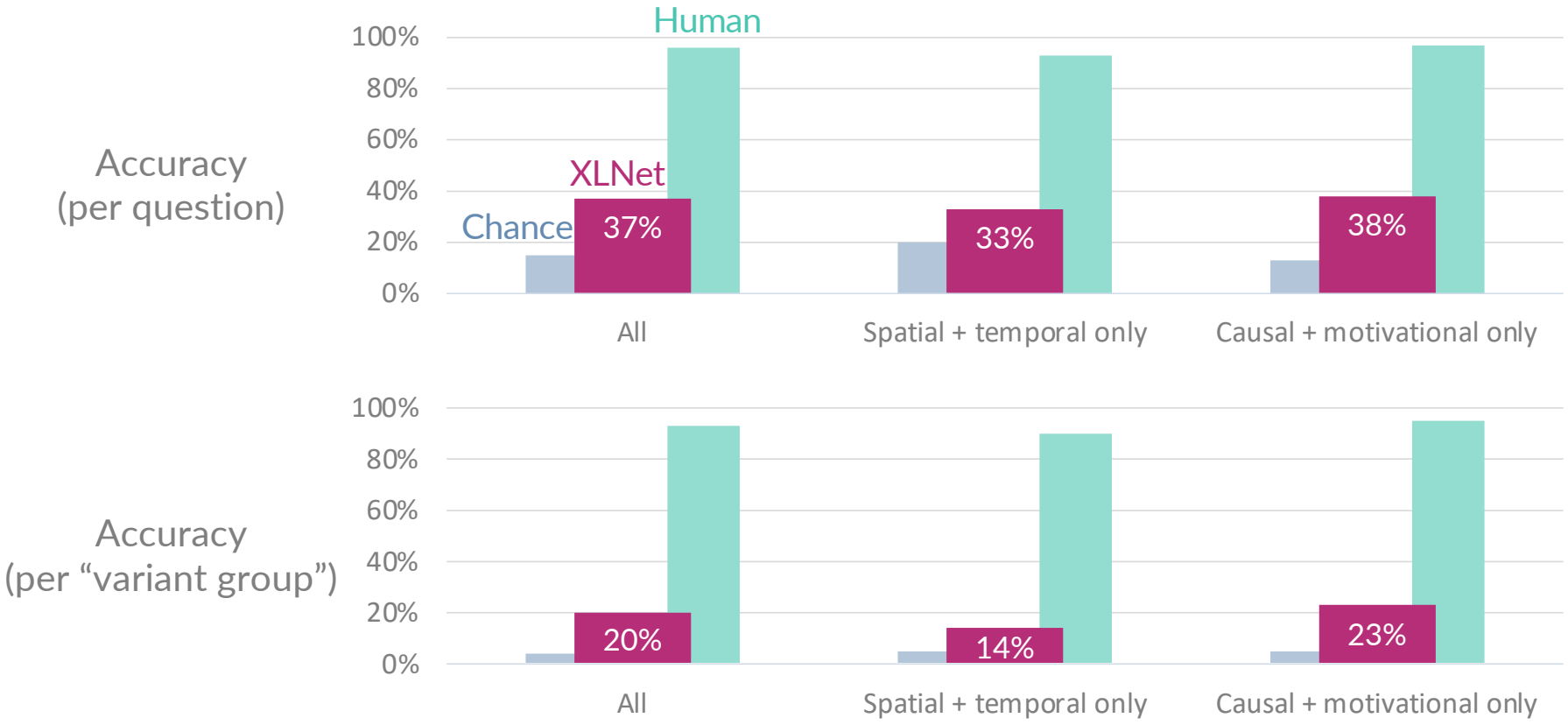
A. The directions the man gave led to where Mr. Green wanted to go.
B. **The directions the man gave led to where the man wanted to go.**
C. The directions Mr. Green gave led to where the man wanted to go.
D. The directions Mr. Green gave led to where Mr. Green wanted to go.

201 questions (91 "variant groups")

* Lai et al., 2017

XLNet*, which performed well on prior data sets (>80%), performs poorly on ToU-based questions.

That's even within the guardrails of multiple-choice.



* Yang et al., 2019

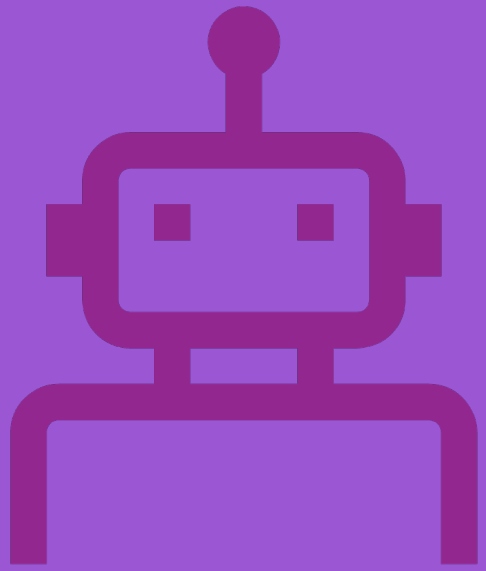
Our ToU test was revealing BUT still a weak test for understanding

- Not just Multiple Choice
- Full Text Response
- Graphical responses -“Draw me a picture”
- Interactive, explanatory dialogs - answer why
- Did understanding improve user’s ability to make better decisions

Machine Understanding is hard.

Let's design benchmarks that acknowledge that and

can move us forward to achieve a more ambitious vision.



Thank you

ELEMENTAL
cognition