



NSF OKN Workshop @ KGC 2023

Thoughts about **interoperability** and **open standards** for knowledge graphs

Dr. Ora Lassila

Principal Technologist
Amazon **Neptune**



Who am I?

Principal Technologist in the Amazon Neptune graph database team

Co-author of

- the original RDF specification
- the seminal paper on the Semantic Web

Recipient of the 1st ISWC “10-year award”

Former W3C Fellow & 7-term elected member of the W3C Advisory Board

Co-author of the NSF “Open Knowledge Network Roadmap” report

Ph.D CS, Helsinki University of Technology

This is not a talk about LLMs...

This is not a talk about LLMs...

(mostly)

Won't get fooled again...

“Meet the new silos... just like the old silos”

Old silos: Single-application -controlled data, at best behind a bespoke API

New silos: Single-purpose knowledge graphs built without interoperability and interlinking in mind

Why do we need interoperability (and what does it mean)?

Common formats enable information interchange

Common query languages enable interworking and free users from "lock-in"

BUT...

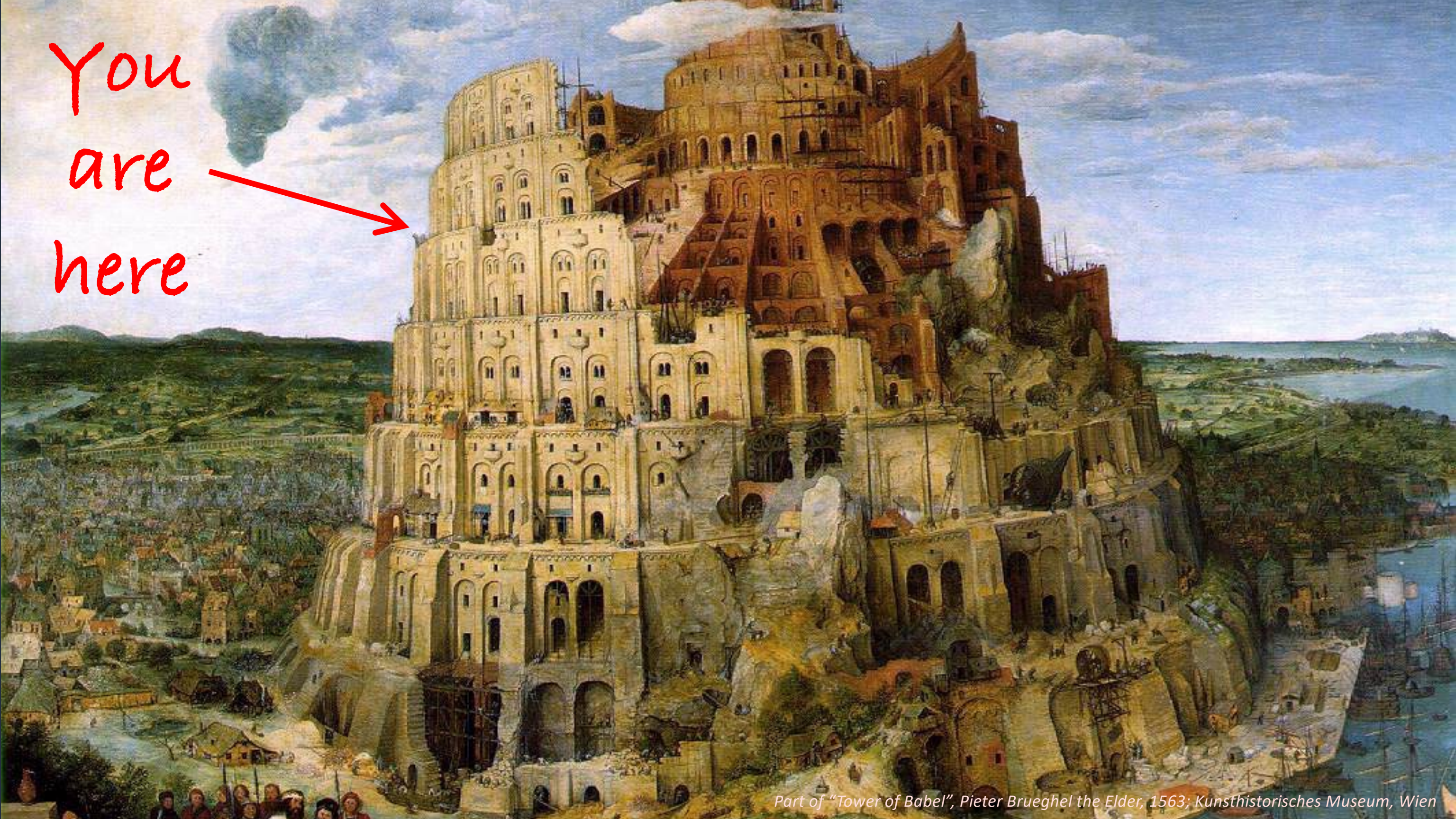
We also need common, shared **semantics**

Special attention should be paid to how we identify things

Standards or technologies not designed for sharing and interchange of semantics should be rejected off-hand

- because they simply just reinforce the old "silo mindset"

You
are
here

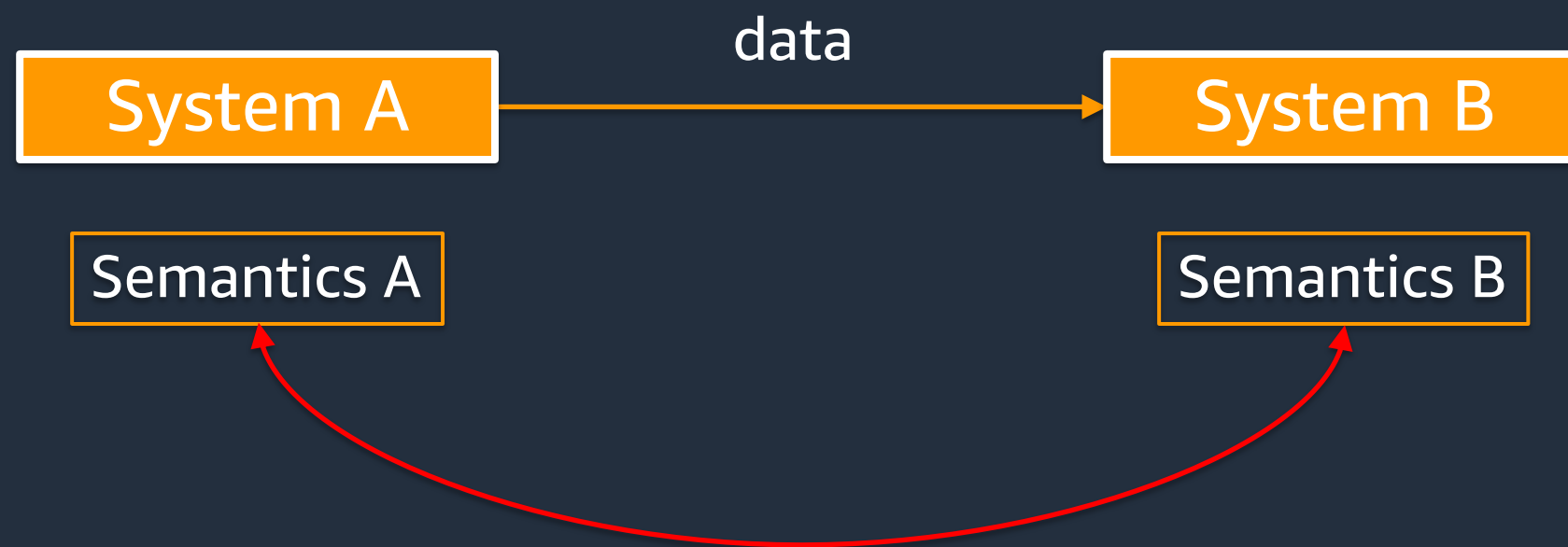


Part of "Tower of Babel", Pieter Bruegel the Elder, 1563; Kunsthistorisches Museum, Wien

This is what we have

Data moves, but not semantics: redefining or “re-articulating” semantics over and over is error-prone and a lot more work

→ technical debt

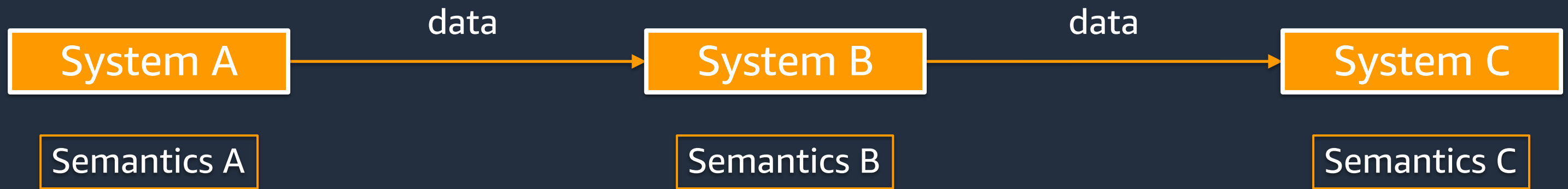


Acting on System A's data in System B requires re-coding A's semantics

No guarantees these are in any way related to one another!

This is what we have

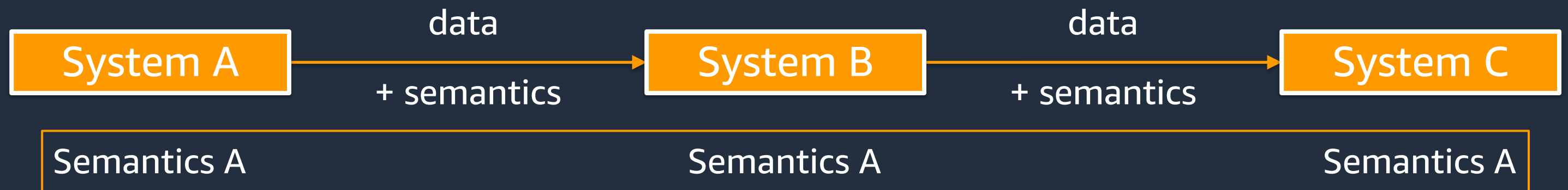
And we keep repeating this...



Remember: accessible data = physical bits + semantics

This is what we should have!

Data moves **with** semantics!



What do we need to get there?

What if acting on data meant you just had to interpret the ontology that defines the semantics of said data?

1. Semantics:

defines how data “behaves” and how software can **interpret** data

2. Ontologies:

provide means of communicating #1

Existing open standards

Modern KGs typically use the **W3C "Semantic Web stack"**

- RDF & OWL, SPARQL, SHACL, R2RML, URI
- mature specifications, work started in 1997...

More recently, a "competing" graph metamodel has appeared: **Labeled Property Graphs (LPG)**

- this is more like a set of *de facto* standards and open source implementations
- LPG databases have been designed more like conventional databases, knowledge interchange has not been a priority

Existing open standards

RDF was specifically designed for interoperability and knowledge interchange

Identifiers: URIs

- globally unique
- suitable for inter-systems linking

ETL & data on-the-wire

- RDF has formally defined semantics for graph merging
- several standardized, interoperable serialization syntaxes
- R2RML for ETL, SHACL for validation

RDF Schema and **OWL** are ontology languages well suited for a broad spectrum of knowledge capture needs

Existing open standards

SPARQL was intended to be the sole means by which an application and a graph database (knowledge graph store) communicate with one another

SPARQL supports **federated queries**

- this makes it easy to build systems that leverage other (external) knowledge graphs and data sources

These two features make it very easy to build distributed, loosely coupled knowledge systems

Existing open standards

One typically wants to base a KG on an ontology, or a set of ontologies

- many public ontologies exist, you do not need to start from scratch
- RDF was designed for extension

Common, useful vocabularies:

- Dublin Core
- SKOS
- PROV-O
- (many others...)

Existing open standards: emerging stuff...

W3C RDF-star

- goal: make it easy to express “statements about statements”
- (current RDF mechanism, reification, is awkward and unpopular)
- brings RDF closer to LPGs (“edge properties”)

ISO GQL

- goal: standard query language for LPGs (loosely based on Cypher)
- also: a schema language for LPGs! (finally)
- worked on by LDBC, feeds ISO

What's next?

The “rift” between RDF and LPGs continues to plague the industry and causes all kinds of confusion

The real issue, however, is this:

- graph as a **logical representation** vs. graph as a **data structure**

So, ostensibly, choosing should not be hard, but the fact remains that neither set of technologies is perfect

A graph is a graph is a graph



The Amazon Neptune team is working to mitigate the rift: **Project OneGraph**

semantic-web-journal.net/system/files/swj3273.pdf

The solution is not easy

- **RDF-star** is a significant step in the right direction

There are benefits "on both sides":

- use all of the good features of RDF (and SPARQL) with LPGs, without having to reinvent them
- no more complaints that RDF does not have "edge properties" *
- Gremlin queries over RDF!

We are working on an implementation that will be available in the future

* Caveat: we still need a semantic theory for these

Enough about KGs, what about LLMs...?

Knowledge graphs can provide “grounding” for LLMs

- curated, trusted sources
- “explainability”
- provenance, lineage

The long span of AI: we are now at a point where we can **really mess things up**

- (again)
- **trustworthiness** of technologies is the key



The screenshot shows the top portion of a news article on The Washington Post website. The header includes the site's logo, a 'Subscribe' button, and a user profile icon. Below the header is a navigation menu with categories like 'Tech', 'Help Desk', 'Future of Transportation', 'Innovations', 'Internet Culture', 'Space', and 'Tech Policy'. The article is categorized under 'INNOVATIONS' and has a main headline: 'ChatGPT invented a sexual harassment scandal and named a real law prof as the accused'. A sub-headline reads: 'The AI chatbot can misrepresent key facts with great flourish, even citing a fake Washington Post article as evidence'. The byline is 'By Pranshu Verma and Will Oremus' and the date is 'April 5, 2023 at 2:07 p.m. EDT'.

Thank you!

Contact:

- ora@amazon.com
- Twitter: [@oralassila](https://twitter.com/oralassila)

